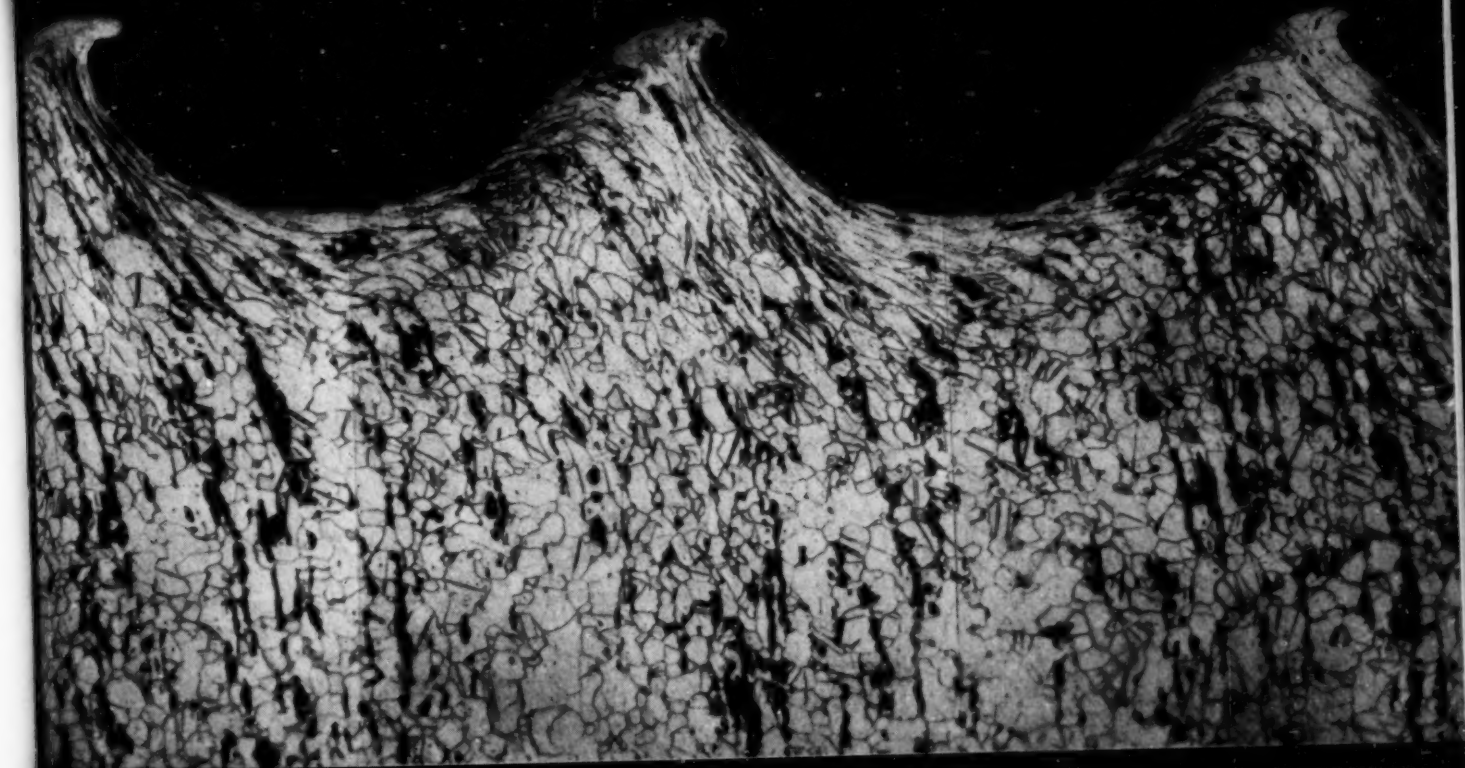
JOURNAL OF APPLIED PHYSICS



JOURNAL OF APPLIED PHYSICS

Vol. 25, No. 5 MAY, 1954

In This Issue

	Page
Circumferential Gap in a Circular Wave Guide Excited by a Dominant Circular-Electric Wave — Leonard S. Sheingold and James E. Storer	545
Impedance of a Top-Loaded Antenna of Arbitrary Length over a Circular Grounded Screen — James R. Wait and Walter J. Surtees	553
Plastic Deformation in Beams under Distributed Dynamic Loads — J. A. Seiler and P. S. Symonds	556
Tube Noise under Large Transit-Time Conditions — S. V. Yadavalli	564
Some Phenomena Associated with Supersonic Liquid Jets — Brian Dunne and Benedict Cassen	569
Transverse Motion of an Electron in a Constant Wave Speed Section of a Linear Accelerator — Ira Jacobs and E. S. Akeley	572
Friction of Clean Metals and Oxides with Special Reference to Titanium — E. S. Machlin and W. R. Yankee	576
On the Band Width of Cavity Antennas — Marshall H. Cohen	582
Multiple Shock Reflection in Corners — Joseph B. Keller	588
Measurement of the Effect of Chlorine Treatment on the Work Function of Titanium and Zirconium — Herbert Malamud and Aaron D. Krumbein	591
Delayed Yield and Strain Rate and Temperature Dependence of Yield Point in Iron — Takeo Yokobori	593
Negentropy and Information in Telecommunications, Writing, and Reading — L. Brillouin	595
Electrolytic Analog Transistor — Harry Letaw, Jr., and John Bardeen	600
Diffusion in Bimetal Vapor-Solid Couples — R. W. Balluffi and L. L. Seigle	607
Stored Energy and Power Flow in Electron Beams — L. R. Walker	615
Information Theory and Knowledge — D. K. C. MacDonald	619
Limiting Pressure on Hydrofoils at Small Submergence Depths — E. V. Laitone	623
Fast Time Analysis of Intermittent Point-to-Plane Corona in Air. III. The Negative Point Trichel Pulse Corona — M. R. Amin	627
Volume and Surface Recombination Rates for Injected Carriers in Germanium — J. P. McKelvey and R. L. Longini	634
Propagation of Microwaves between a Parallel Pair of Doubly Curved Conducting Surfaces — K. S. Kunz	642
Errors in Power Spectra Due to Finite Sample — Lee M. Spetner	653
Approximate Emissivity Calculations for Polyatomic Molecules. I. CO_2 — S. S. Penner	660
Cross Section of Pulled Crystals — Robert G. Pohl	668
Some New Magneto-Mechanical Torsion Experiments — A. W. Cochardt	670
Letters to the Editor:	
Electron Bunches of Short Time Duration — E. W. Ernst and H. VonFoerster	674
Effect Produced by Oblique Impact of a Cylinder on a Thin Target — William A. Allen, Joe M. Mapes, and Wesley G. Wilson	675
New Silicon *p-n* Junction Photocell for Converting Solar Radiation into Electrical Power — D. M. Chapin, C. S. Fuller, and G. L. Pearson	676
Dissected Amplifiers Using Negative Resistance — W. Shockley and W. P. Mason	677
On the Interaction of Two Plane Shocks Facing in the Same Direction — Otto Laporte and E. B. Turner	678
High-Speed Magnetic Pulsing of Ferrites — R. Conway LeCraw	678
Crossed Electron Beam Technique for Measuring Space-Charge Effects in Beams — Paul D. Coleman, Murray D. Sirkis, and Irving Kaufman	679
Combined Slow-Wave Circuit and Beam Focusing Structure — Philip Rice	680
Announcements	681
Index to Advertisers	xxii

ELMER HUTCHISSON, *Editor*

EARLE C. GREGG, *Assistant Editor*

Board of Editors

MAURICE A. BIOT
RAY E. BOLZ
WALLACE J. ECKERT
ROBERT D. HEIDENREICH
LOUIS A. PIPES
ROMAN SMOLUCHOWSKI
W. R. SMYTHE
ARTHUR V. TOBOLSKY
ALVIN M. WEINBERG

Publication of the American Institute of Physics

OFFICERS

FREDERICK SEITZ, *Chairman*
GEORGE B. PEGRAM, *Treasurer*

GOVERNING BOARD

ROBERT F. BACHER
J. W. BUCHTA
KARL K. DARROW
S. A. GOUDSMIT
DEANE B. JUDD
HUGH S. KNOWLES
C. C. LAURITSEN
W. H. MARKWOOD, JR.
WILLIAM F. MEGGERS
PHILIP M. MORSE
BRIAN O'BRIEN
HARRY F. OLSON
GEORGE B. PEGRAM
ERIC RODGERS
RALPH A. SAWYER
FREDERICK SEITZ
WILLIAM SHOCKLEY
RICHARD M. SUTTON
J. H. VAN VLECK
MARK W. ZEMANSKY

Consultant on Printing
MELVIN LOOS

ADMINISTRATIVE STAFF

HENRY A. BARTON, *Director*
WALLACE WATERFALL, *Executive Secretary*
THEODORE VORBURGER, *Advertising Manager*
RUTH F. BRYANS, *Publication Manager*
EDITH I. NEFTEL *Circulation Manager*
KATHRYN SETZE *Assistant Treasurer*

The JOURNAL OF APPLIED PHYSICS, published monthly at Prince and Lemon Sts., Lancaster, Pa., is devoted to physics in its role as the science basic to other natural sciences and to the arts and industries. Previous to 1937 this publication was known as PHYSICS. The JOURNAL OF APPLIED PHYSICS publishes editorials and reviews, as well as technical papers of applied physics. Articles appearing in it are indexed regularly, according to their field, by Physics Abstracts, Chemical Abstracts, Engineering Index, Inc., Industrial Arts Index, etc.

Manuscripts should be submitted to Elmer Hutchisson, Editor, Case Institute of Technology, Cleveland 6, Ohio. Unless otherwise stated, submission of a manuscript is a representation that it has been neither copyrighted, published, nor submitted for publication elsewhere. The authors' institutions are requested to pay a **publication charge** of $8 per page which, if honored, entitles them to 100 reprints without further charge. Instructions will be sent with galley proofs.

Subscriptions, renewals, and orders for back numbers should be addressed to the American Institute of Physics, 57 East 55 Street, New York 22, New York.

Subscription Price	U.S. and Canada	Elsewhere
To members of the American Institute of Physics and Affiliated Societies	$10.00	$11.00
To all others	12.00	13.00

Back Number Prices

Yearly back number rate when complete year is available: $14.00. Single copies: $1.25 each.

Changes of address should be sent to the American Institute of Physics; **corrected proofs** should be sent to the Publication Manager, American Institute of Physics. **Advertising rates** supplied on request. Orders, advertising copy, and cuts should be sent to the American Institute of Physics.

Entered as second class matter January 22, 1937, at the Post Office at Lancaster, Pennsylvania, under the Act of March 3, 1879. Accepted for mailing at the special rate of postage provided for in paragraph (d-2), section 34.40, P. L. & R. of 1948, authorized May 2, 1932.

Journal
of
Applied Physics

Volume 25, Number 5 May, 1954

Circumferential Gap in a Circular Wave Guide Excited by a Dominant Circular-Electric Wave

LEONARD S. SHEINGOLD* AND JAMES E. STORER
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received July 23, 1953)

The scattering matrix of a 360-degree circumferential slot in an infinitely long circular wave guide with an incident dominant circular-electric wave is obtained by a variational principle. The theory which should by very good for small gaps is shown to be in good agreement with the obtained experimental results.

1. INTRODUCTION

CONSIDER a circumferential gap in an infinitely long, perfectly conducting circular wave guide as shown in Fig. 1. The gap can be considered to be a region of free-space coupling two semi-infinite circular wave guides. Assuming that the gap width is small compared to the guide wavelength, the basic problem is to determine the effect of the gap on an incident TE_{01} wave.

The energy in the TE_{01} wave incident on the gap is partially reflected and partially transmitted. Some of the energy is coupled through the gap to the free space surrounding the wave guide. Since the power relationships are the essential quantities of interest, the problem is formulated in terms of the scattering matrix of a wave-guide junction.

Intuitively, the small gap should have little effect on the propagation of an incident circular-electric wave. The circular-electric modes have the unique property that the wall surface currents are purely transverse. These transverse currents do not tend to charge the metallic boundary of the circumferential gap; hence little energy is expected to radiate into space. On the other hand, all TM modes are characterized by solely longitudinal currents on the guide wall, and all TE modes other than the circular-electric type have longitudinal components of current. It can be concluded that the modal energy of all extraneous modes will be coupled strongly to the surrounding medium.

The problem is formulated in terms of an integral equation for the axial component of the magnetic field on the gap surface. A variational principle is applied and a parameter Δ is introduced representing the stationary parameter in the variational formulation of the integral equation. For simplification, even and odd excitations are used and the elements of the scattering matrix are expressed in terms of the appropriate Δ's. The restriction imposed by the theory is that the gap width must be small compared with wavelength.

2. FORMULATION OF THE PROBLEM

Since the following analysis pertains to circularly symmetrical wave-guide modes, the electromagnetic field components can be derived from a scalar function which is independent of the ϕ coordinate. Assuming a time dependence $\exp -i\omega t$, the circular-electric modes are obtained by solving the equation

$$LE_\phi(r,z)=0, \tag{1}$$

where the operator L acting on the ϕ components of the field is given by

$$L=\frac{\partial^2}{\partial r^2}+\frac{1}{r}\frac{\partial}{\partial r}+\frac{\partial^2}{\partial z^2}-\frac{1}{r^2}+k^2. \tag{2}$$

FIG. 1. A circumferential gap in an infinitely long circular wave guide.

* Now with the Boston Engineering Laboratory of the Sylvania Electric Products, Inc., Boston, Massachusetts. The research reported in this document was made possible through support extended Cruft Laboratory, Harvard University, jointly by the Navy Department (U. S. Office of Naval Research), the Signal Corps of the U. S. Army, and the U. S. Air Force.

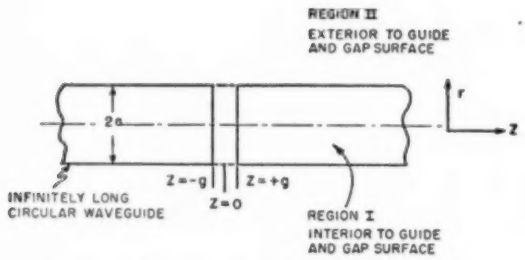


FIG. 2. Orientation of cylindrical coordinate system.

The other field components can be expressed in terms of E_ϕ by

$$H_r(r,z)=\frac{i}{\omega\mu}\frac{\partial}{\partial z}E_\phi(r,z),$$

$$H_z(r,z)=\frac{-i}{\omega\mu}\frac{1}{r}\frac{\partial}{\partial r}rE_\phi(r,z), \tag{3}$$

$$E_r=E_z=H_\phi=0.$$

The walls of the cylinder are assumed to be perfectly conducting; therefore, the boundary conditions at the wall require that the tangential components of the electric field vanish. Hence, the boundary condition is

$$E_\phi(r,z)\big]_{r=a}=0. \tag{4}$$

A circumferential gap, of width $2g$, centered at $z=0$ is introduced in the infinite guide as indicated in Fig. 2. The region for $r<a$ will be designated by superscript I and the region exterior to the guide, $r>a$, by superscript II.

In region II ($r>a$) the field can be expressed in terms of a Green's function and the field distribution on the surface of the gap. The Green's function is defined by the equation

$$LG^{II}(r,r',z-z')=-\frac{\delta(r-r')\delta(z-z')}{r'} \quad r>a$$

$$G^{II}(r,r',z-z')\big]_{r=a}=0. \tag{5}$$

By making use of Green's second theorem, an integral representation of the external field can be found in the form

$$E_\phi^{II}(r,z)=\int_{-g}^{g}\left[\frac{1}{r'}\frac{\partial}{\partial r'}r'G^{II}(r,r',z-z')\right]_{r'=a}\xi(z')dz', \tag{6}$$

where

$$\xi(z')=[E_\phi(r,z')]_{r=a}.$$

While an explicit form for this Green's function can be obtained if desired, it is not necessary. Actually, it is simpler to determine the function occurring in the integral of (6). It is apparent from (6) that the integrand must satisfy the two conditions

$$L\left[\frac{1}{r'}\frac{\partial}{\partial r'}r'G^{II}(r,r',z-z')\right]_{r'=a}=0 \quad r>a \tag{7}$$

$$\left[\frac{1}{r'}\frac{\partial}{\partial r'}r'G^{II}(r,r',z-z')\right]_{r'=a,\ r=a}=\delta(z-z') \tag{8}$$

as well as the Sommerfeld radiation condition. By the use of the Fourier transform one obtains an expression of the form[1]

$$\left[\frac{1}{r}\frac{\partial}{\partial r'}r'G^{II}(r,r',z-z')\right]_{r'=a}=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{i\zeta(z-z')}\frac{H_1^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}r]}{H_1^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a]}d\zeta, \tag{9}$$

where

$$(k^2-\zeta^2)^{\frac{1}{2}}=i(\zeta^2-k^2)^{\frac{1}{2}} \quad \text{for} \quad \zeta>k.$$

It is immediately apparent that such a choice satisfies (7) and (8) as well as the radiation conditions. Since the Green's function is unique, (9) is, therefore, the desired function.

A similar analysis can also be carried out for the region I ($r>a$). Since there are fields propagating inside the pipe, the resulting expressions for the fields include an additional term representing the incident wave. In this case one obtains

$$E_\phi^{I}(r,z)=E_{\phi\,\text{Inc.}}^{I}(r,z)-\int_{-g}^{g}\left[\frac{1}{r'}\frac{\partial}{\partial r'}r'G^{I}(r,r',z-z')\right]_{r'=a}\xi(z')dz'. \tag{10}$$

Performing an analysis similar to that for region II, it is possible to show that

$$-\left[\frac{1}{r'}\frac{\partial}{\partial r'}r'G^{I}(r,r',z-z')\right]_{r'=a}=\frac{1}{2\pi}\int_{C}e^{i\zeta(z-z')}\frac{J_1[(k^2-\zeta^2)^{\frac{1}{2}}r]}{J_1[(k^2-\zeta^2)^{\frac{1}{2}}a]}d\zeta. \tag{11}$$

Assuming only a single mode propagates in the pipe, the Bessel function in the denominator of (11) vanishes at two places ($\pm\gamma_1$) on the real axis, and at an infinite set of points ($\pm i\gamma_n$) along the imaginary axis. The contour C used in (11) is that indicated in Fig. 3. The contour is chosen to pass around the singularities in the manner shown so that the Green's function represents only waves propagating away from the junction.

Thus, the entire electromagnetic field has been expressed in terms of the field on the surface of the gap. It remains now to determine integral equations for

[1] C. Papas, Radiation from a Transverse Slot in an Infinite Cylinder, Cruft Laboratory Tech. Report, No. 61, 1950.

these fields. This can be done by matching the magnetic fields across the gap.

In particular, from (9) the magnetic field in the gap surface can be shown (except for dimensional constants) to be

$$H_z^{\mathrm{II}}(a,z)\simeq\left[\frac{1}{r}\frac{\partial}{\partial r}rE_\phi^{\mathrm{II}}(r,z)\right]_{r=a}$$

$$=\int_{-g}^{g}K^{\mathrm{II}}(z-z')\xi(z')dz', \quad (12)$$

where

$$K^{\mathrm{II}}(z-z')=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{i\zeta(z-z')}H(\zeta)d\zeta \quad (13)$$

and

$$H(\zeta)=\frac{H_0^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a]}{H_1^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a]}(k^2-\zeta^2)^{\frac{1}{2}}. \quad (14)$$

Similarly, from (10)

$$H_z^{\mathrm{I}}(a,z)\simeq\left[\frac{1}{r}\frac{\partial}{\partial r}rE_\phi^{\mathrm{I}}(r,z)\right]_{r=a}=\left[\frac{1}{r}\frac{\partial}{\partial r}rE_{\phi\ \mathrm{Inc.}}^{\mathrm{I}}(r,z)\right]_{r=a}$$

$$+\int_{-g}^{g}K^{\mathrm{I}}(z-z')\xi(z')dz', \quad (15)$$

where

$$K^{\mathrm{I}}(z-z')=\frac{1}{2\pi}\int_{C}e^{i\zeta(z-z')}J(\zeta)d\zeta \quad (16)$$

and

$$J(\zeta)=\frac{J_0[(k^2-\zeta^2)^{\frac{1}{2}}a]}{J_1[(k^2-\zeta^2)^{\frac{1}{2}}a]}(k^2-\zeta^2)^{\frac{1}{2}}. \quad (17)$$

Now

$$H_z^{\mathrm{II}}(a,z)=H_z^{\mathrm{I}}(a,z) \quad -g<z<g,$$

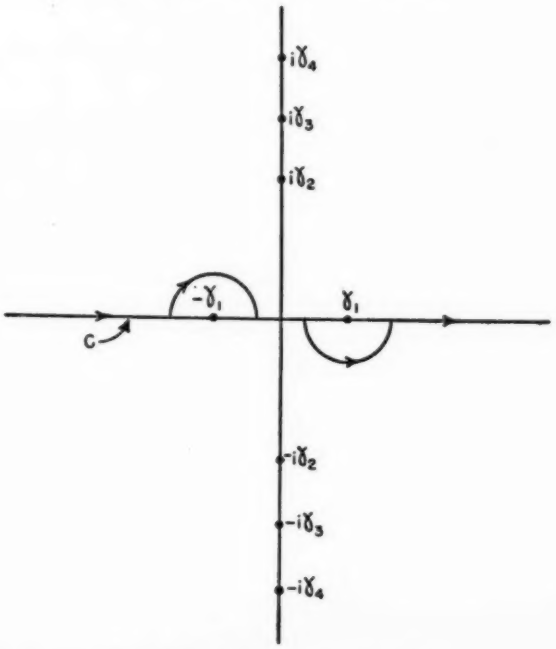


FIG. 3. Contour C in complex plane.

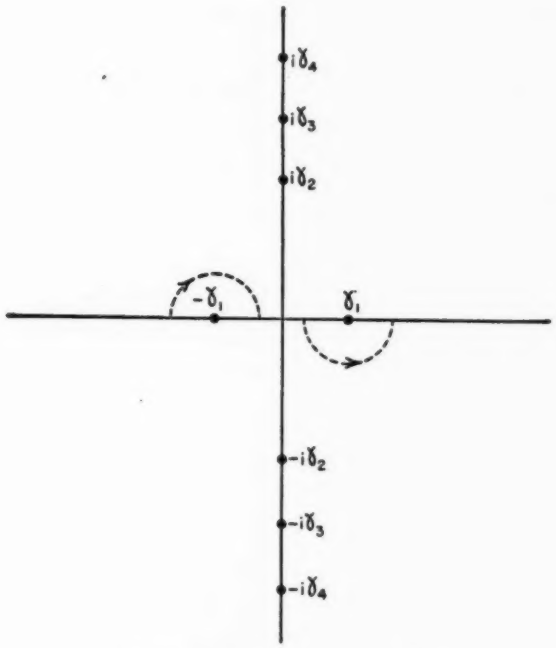


FIG. 4. Contour $\bar{C}$ in complex plane.

and, therefore, one obtains the integral equation

$$\left[\frac{1}{r}\frac{\partial}{\partial r}rE_{\phi}^{\mathrm{I}\ \mathrm{Inc.}}(r,z)\right]_{r=a}=\int_{-g}^{g}[K^{\mathrm{II}}(z-z')-K^{\mathrm{I}}(z-z')]$$

$$\times\xi(z')dz' \quad -g<z<g. \quad (18)$$

3. VARIATIONAL FORMULATION

It is now convenient to separate $K^{\mathrm{I}}(z-z')$ into two terms, as follows:

$$K^{\mathrm{I}}(z-z')=\bar{K}^{\mathrm{I}}(z-z')+\frac{1}{2\pi}\int_{C}e^{i\zeta(z-z')}J(\zeta)d\zeta, \quad (19)$$

where

$$\bar{K}^{\mathrm{I}}(z-z')=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{i\zeta(z-z')}J(\zeta)d\zeta, \quad (20)$$

i.e., integrals along the real axis and similar in form to those of (13). The remaining integrals along the contour $\bar{C}$ indicated in Fig. 4 can be evaluated by means of residues.

The integral equation then assumes the form

$$\frac{i\gamma_1a^3}{\rho_1^2}\left[\frac{1}{r}\frac{\partial}{\partial r}rE_{\phi}^{\mathrm{I}\ \mathrm{Inc.}}(r,z)\right]_{r=a}+\int_{-g}^{g}\cos\gamma_1(z-z')\xi(z')dz'$$

$$=\int_{-g}^{g}K(z-z')\xi(z')dz', \quad (21)$$

where

$$K(z-z')=\frac{\gamma_1a^3}{i\rho_1^2}[K^{\mathrm{II}}(z-z')-\bar{K}^{\mathrm{I}}(z-z')] \quad (22)$$

and, the constants ρ_1, γ_1 are defined by

$$J_1(\rho_1)=0; \quad \gamma_1=\left[k^2-\left(\frac{\rho_1}{a}\right)^2\right]^{\frac{1}{2}}.$$

The fields incident in the wave guide can be expressed in the form

$$E_\phi{}^{\mathrm{I}}{}_{\mathrm{Inc.}}(r,z)=\frac{\rho_1}{i\gamma_1 a^2}\frac{J_1\left(\rho_1\frac{r}{a}\right)}{J_0(\rho_1)}\alpha(z). \tag{23}$$

Their z dependence, $\alpha(z)$, is, in general, of the form

$$\alpha(z)=Ae^{i\gamma_1 z}+Be^{-i\gamma_1 z},$$

i.e., fields incident on the gap from either the right or left. It is convenient to specialize this into even or odd incident fields, i.e.,

$$\alpha_e(z)=\cos\gamma_1 z \quad \text{even}$$
$$\alpha_0(z)=\sin\gamma_1 z \quad \text{odd}.$$

For the even case, $\xi_e(z)=\xi_e(-z)$ and hence,

$$\int_{-g}^{g}\cos\gamma_1(z-z')\xi_e(z')dz'=\alpha_e(z)\int_{-g}^{g}\alpha_e(z')\xi_e(z')dz'.$$

Similarly, for the case of odd excitation $\xi_0(z)=-\xi_0\times(-z)$ and hence,

$$\int_{-g}^{g}\cos\gamma_1(z-z')\xi_0(z')dz'=\alpha_0(z)\int_{-g}^{g}\alpha_0(z')\xi_0(z')dz'.$$

Therefore, for either *even* or *odd* excitation, (23) can be inserted in the integral equation (21) which then becomes:

$$\alpha(z)\left[1+\int_{-g}^{g}\alpha(z')\xi(z')dz'\right]=\int_{-g}^{g}K(z-z')\xi(z')dz'. \tag{24}$$

Multiplying (24) by $\xi(z)$, integrating from $-g$ to g and dividing by

$$\left[\int_{-g}^{g}\alpha(z)\xi(z)dz\right]^2$$

yields the variational principle

$$\Delta=\frac{1+\int_{-g}^{g}\alpha(z)\xi(z)dz}{\int_{-g}^{g}\alpha(z)\xi(z)dz}=\frac{\int_{-g}^{g}\int_{-g}^{g}K(z-z')\xi(z)\xi(z')dzdz'}{\left[\int_{-g}^{g}\alpha(z)\xi(z)dz\right]^2}. \tag{25}$$

Equation (25) can be shown to have all the usual variational properties, i.e., Δ is stationary with respect to small changes of $\xi(z)$ about its correct value and Δ is independent of the over-all amplitude of $\xi(z)$.

Thus, it is possible to insert an approximate value for $\xi(z)$ in (25) and obtain relatively more accurate values of Δ. As will be seen later, the various Δ's can be related to the scattering matrix of the junction.

Referring to Eqs. (13), (14), (17), (20), and (24), it is seen that $K(z-z')$ can be expressed in the form

$$K(z-z')=\int_{-\infty}^{\infty}e^{i\zeta(z-z')}\psi(\zeta)d\zeta \tag{26}$$

where, using the Wronskian for Bessel functions,

$$\psi(\zeta)=\frac{\gamma_1 a^2}{\pi^2\rho_1{}^2}\frac{1}{J_1[(k^2-\zeta^2)^{\frac{1}{2}}a]H_1{}^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a]}. \tag{27}$$

The variational principles can then be written in the form

$$\Delta=\int_{-\infty}^{\infty}\Phi^2(\zeta)\psi(\zeta)d\zeta, \tag{28}$$

where

$$\Phi_e(\zeta)=\frac{\int_{-g}^{g}\cos\zeta z\xi_e(z)dz}{\int_{-g}^{g}\cos\gamma_1 z\xi_e(z)dz},$$
$$\Phi_0(\zeta)=\frac{\int_{-g}^{g}\sin\zeta z\xi_0(z)dz}{\int_{-g}^{g}\sin\gamma_1 z\xi_0(z)dz}. \tag{29}$$

It remains now to choose trial functions for $\xi(z)$ and insert them into (28).

Consider the following choices

$$\tilde{\xi}_e(z)=(g^2-z^2)^{\frac{1}{2}}\tilde{\xi}_0(z)=z(g^2-z^2)^{\frac{1}{2}}. \tag{30}$$

It is apparent that these functions satisfy the symmetry conditions and have the correct static dependence on the edge of the gap. For small gaps they should, therefore, be good approximations to the correct fields. Inserting these into (29) yields

$$\Phi_e(\zeta)=\frac{\gamma_1 J_1(\zeta g)}{\zeta J_1(\gamma_1 g)}$$
$$\Phi_0(\zeta)=\frac{[2/(\zeta g)^2]J_1(\zeta g)-[J_0(\zeta g)/\zeta g)]}{[2J_1(\gamma_1 g)/(\gamma_1 g)^2]-[J_0(\gamma_1 g)/(\gamma_1 g)]}. \tag{31}$$

Inserting (31) into (28), it can be shown (see Appendix) that

$$\Delta_e\doteq\frac{\gamma_1 a}{\rho_1{}^2}\left\{\frac{(ka)^2}{4}+\frac{i4}{\pi}\left(\frac{a}{g}\right)^2\right\}$$
$$\Delta_0\doteq\frac{1}{\gamma_1 a\rho_1{}^2}\left\{\frac{(ka)^4}{16}+\frac{i32}{\pi}\left(\frac{a}{g}\right)^4\right\}.$$

It is to be noted that these results are only valid for small gaps, i.e., $g^2/\lambda^2 \ll 1$. Results for larger gaps would be difficult to obtain theoretically because of the lack of satisfactory trial functions for $\xi(z)$. Moreover, it is virtually impossible to make satisfactory experimental measurements for large gaps since the interaction of the gap with the "free space" region outside the guide is of prime importance and a "free space" region is difficult to duplicate experimentally.

4. TRANSMISSION-LINE REPRESENTATION

It is now desirable to express the scattering matrix elements in terms of Δ_e and Δ_0. This can be accomplished by examining the fields inside the wave guide at a terminal plane located a great distance from the gap discontinuity. The evaluation of the scattering (or impedance) matrix requires the knowledge of the explicit asymptotic expression of the Green's function of Eq. (11). This expression is obtained in the Appendix. Substitution of the asymptotic form into Eq. (10) yields

$$E_\phi^{\mathrm{I}}(r,z)=E_{\phi\,\mathrm{Inc.}}^{\mathrm{I}}(r,z)+\frac{i\rho_1}{\gamma_1 a^2}\frac{J_1\left(\rho_1\frac{r}{a}\right)}{J_0(\rho_1)}e^{-i\gamma_1 z}\int_{-g}^{g}e^{i\gamma_1 z'}\xi(z')dz' \tag{33}$$

for the electric field in the far zone.

It then follows that the z dependence is of the form

$$E_\phi^{\mathrm{I}}(r,z)\sim\alpha(z)+e^{-i\gamma_1 z}\int_{-g}^{g}e^{i\gamma_1 z'}\xi(z')dz'. \tag{34}$$

Now for the even case ($\alpha(z)=\cos\gamma_1 z$) this expression becomes

$$E_\phi^{\mathrm{I}}(r,z)\sim e^{i\gamma_1 z}+\left[1+2\int_{-g}^{g}\cos\gamma_1 z'\xi(z')dz'\right]e^{-i\gamma_1 z} \tag{35}$$

and the even reflection coefficient Γ_e can be defined as

$$\Gamma_e=1+2\int_{-g}^{g}\cos\gamma_1 z'\xi(z')dz'. \tag{36}$$

Similarly, for odd excitation $\alpha(z)=\sin\gamma_1 z$, the odd reflection coefficient Γ_0 can be expressed as

$$\Gamma_0=-1-2\int_{-g}^{g}\sin\gamma_1 z'\xi(z')dz'. \tag{37}$$

The integrals of Eqs. (36) and (37) can be expressed in terms of Δ from Eq. (25), resulting in

$$\Gamma_e=\frac{\Delta_e-1}{\Delta_e+1},\qquad \Gamma_0=-\frac{\Delta_0-1}{\Delta_0+1}. \tag{38}$$

The elements of the scattering matrix are then given by

$$S_{11}=S_{22}=\frac{\Gamma_e+\Gamma_0}{2}=\frac{1}{1+\Delta_0}-\frac{1}{1+\Delta_e},\qquad S_{12}=\frac{\Gamma_e-\Gamma_0}{2}=1-\frac{1}{1+\Delta_0}-\frac{1}{1+\Delta_e}. \tag{39}$$

5. EXPERIMENTAL AND THEORETICAL RESULTS

Measurements of the characteristics of the circumferential gap were performed in a 5.755-in. i.d. brass tubing at a free-space wavelength of 10.000 cm. Deschamps' method was used to determine the elements of the scattering matrix of the circumferential gap. In comparing the theoretical and experimental results it was convenient to examine separately the absolute values and arguments of the scattering matrix elements. This method of presentation of data was chosen because the experimental values of amplitude and phase were determined separately by Deschamps' method.[2]

The absolute values of S_{11}, S_{22}, and S_{12} are then given by

$$|S_{11}|=|S_{22}|=\frac{|\Delta_e-\Delta_0|}{|1+\Delta_0|\,|1+\Delta_e|},\qquad |S_{12}|=\frac{|\Delta_0\Delta_e-1|}{|1+\Delta_0|\,|1+\Delta_e|}, \tag{40}$$

and their arguments are found to be

$$\mathrm{Arg}S_{11}=\mathrm{Arg}S_{22}=\mathrm{Arg}(\Delta_e-\Delta_0)-\mathrm{Arg}(1+\Delta_0)-\mathrm{Arg}(1+\Delta_e) \tag{41}$$

$$\mathrm{Arg}S_{12}=\mathrm{Arg}(\Delta_e\Delta_0-1)-\mathrm{Arg}(1+\Delta_0)-\mathrm{Arg}(1+\Delta_e).$$

The experimental values are plotted on the same figure as the theoretically determined curves in Figs. 5 and 6.† It is observed that the experimental value of $|S_{11}|$ and $|S_{12}|$ are in excellent agreement with the theory for values of $2g$ up to 1.000 in. corresponding to $g/\lambda=0.127$. It is recalled that the theory is only claimed to be valid for values of $(g/\lambda)^2\ll 1$. As seen from the phase characteristics of S_{11}, the experimental values lie slightly above the theoretical curve for small values of $2g$. It is observed in Fig. 6 that there is again excellent agreement between experiment and theory for values of $\mathrm{Arg}S_{12}$ such that $2g\leq 1.000$.

---

[2] Storer, Sheingold, and Stein, Proc. Inst. Radio Engrs. **41**, 1004–1013 (1953).

† A detailed description of the experimental techniques as well as a more complete discussion of the experimental results is given in Cruft Laboratory Technical Report No. 167. It should also be noted that in calculating the theoretical values of the scattering matrix elements from Eq. (32) $-j$ is substituted for $+i$ to conform to the convention established in measurement.

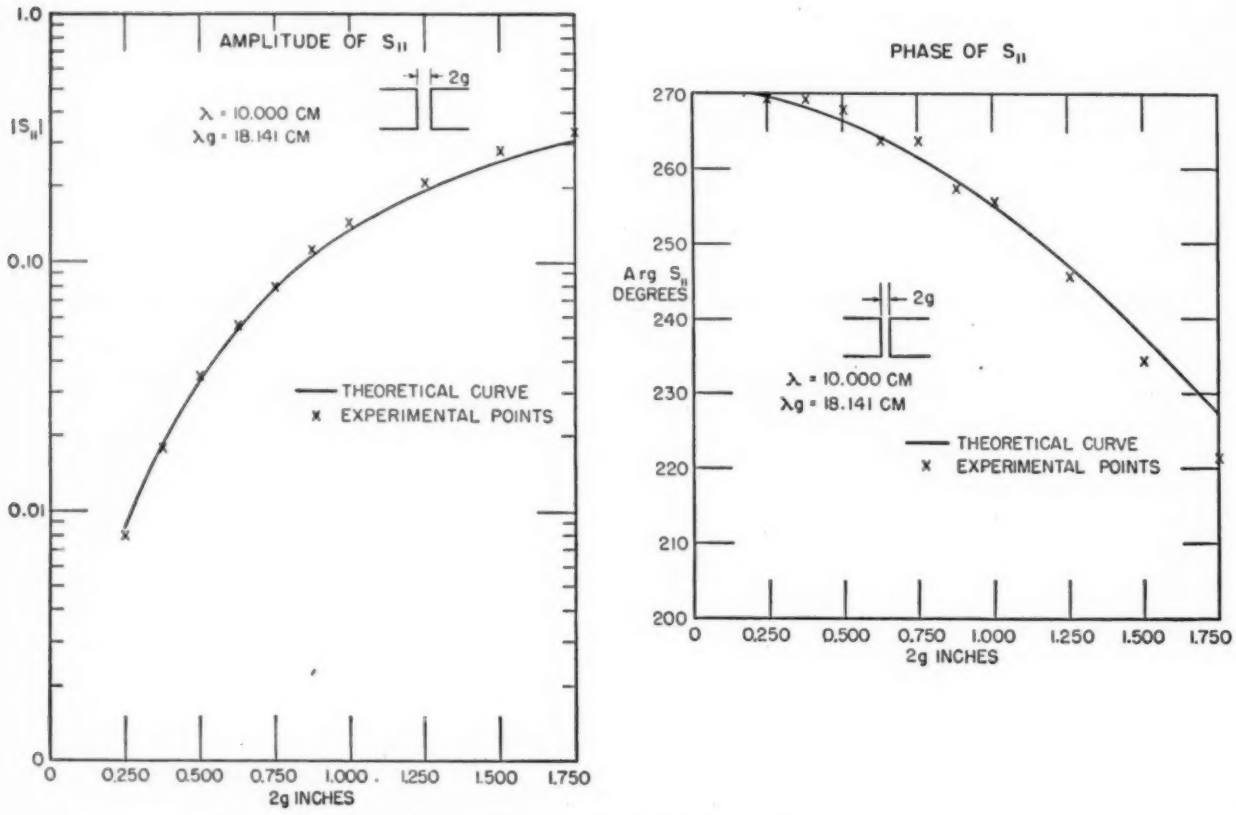


FIG. 5. Measured and calculated values of S_{11}.

APPENDIX A. EVALUATION OF THE VARIATIONAL PARAMETERS Δ_e AND Δ_0

The determination of the elements of the scattering matrix of a circumferential gap requires the evaluation of

$$\Delta_e=\int_{-\infty}^{\infty}\Phi_e^2(\zeta)\psi(\zeta)d\zeta$$
$$\Delta_0=\int_{-\infty}^{\infty}\Phi_0^2(\zeta)\psi(\zeta)d\zeta. \tag{A1}$$

Since the assumed electric field distribution in the gap is a good approximation only for small values of $\gamma_1 g$, the parameters $\Phi_e(\zeta)$ and $\Phi_0(\zeta)$ can be expressed in the approximate form

$$\Phi_e(\zeta)=\frac{\gamma_1 J_1(\zeta g)}{\zeta J_1(\gamma_1 g)}\doteq\frac{2J_1(\zeta g)}{\zeta g}\quad \gamma_1 g\ll 1 \tag{A2}$$

$$\Phi_0(\zeta)=\frac{[2/(\zeta g)^2]J_1(\zeta g)-[J_0(\zeta g)/\zeta g]}{[2J_1(\gamma_1 g)/(\gamma_1 g)^2]-[J_0(\gamma_1 g)/\gamma_1 g]}$$
$$\doteq\frac{8}{\gamma_1 g}\left\{\frac{2}{(\zeta g)^2}J_1(\zeta g)-\frac{J_0(\zeta g)}{\zeta g}\right\}\quad \gamma_1 g\ll 1$$

and $\psi(\zeta)$ is given by

$$\psi(\zeta)=\frac{\gamma_1 a^2}{\pi^2\rho_1^2 J_1[(k^2-\zeta^2)^{\frac{1}{2}}a]H_1^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a]}. \tag{A3}$$

It is convenient to determine separately the real and imaginary parts of Δ_e and Δ_0. Proceeding to evaluate R.P.Δ_e one observes that the integrand is real only for $\zeta<k$. Therefore, assuming $kg\ll 1$ as well as $\gamma_1 g\ll 1$, so that $J_1(\zeta g)\doteq\zeta g/2$ for $0<\zeta<k$, it follows that

$$\text{R.P.}\Delta_e=\frac{2\gamma_1 a^2}{\pi^2\rho_1^2}\int_0^k\frac{d\zeta}{J_1^2[(k^2-\zeta^2)^{\frac{1}{2}}a]+Y_1^2[(k^2-\zeta^2)^{\frac{1}{2}}a]}$$
$$=\frac{2\gamma_1 a^2 j}{\pi^2\rho_1^2}\int_0^1\frac{dt}{|H_1^{(1)}[(1-t^2)^{\frac{1}{2}}ka]|^2}. \tag{A4}$$

For $t\ll 1$, the denominator is large, and hence there is little contribution. On the other hand for t near zero, and with ka equal to 4.592, the Hankel function is already well approximated by its asymptotic form

$$|H_1^{(1)}(1-t^2)^{\frac{1}{2}}ka|^2$$
$$\simeq\frac{2}{\pi ka(1-t^2)^{\frac{1}{2}}}\left[1+\frac{3}{8(1-t^2)(ka)^2}\right]. \tag{A5}$$

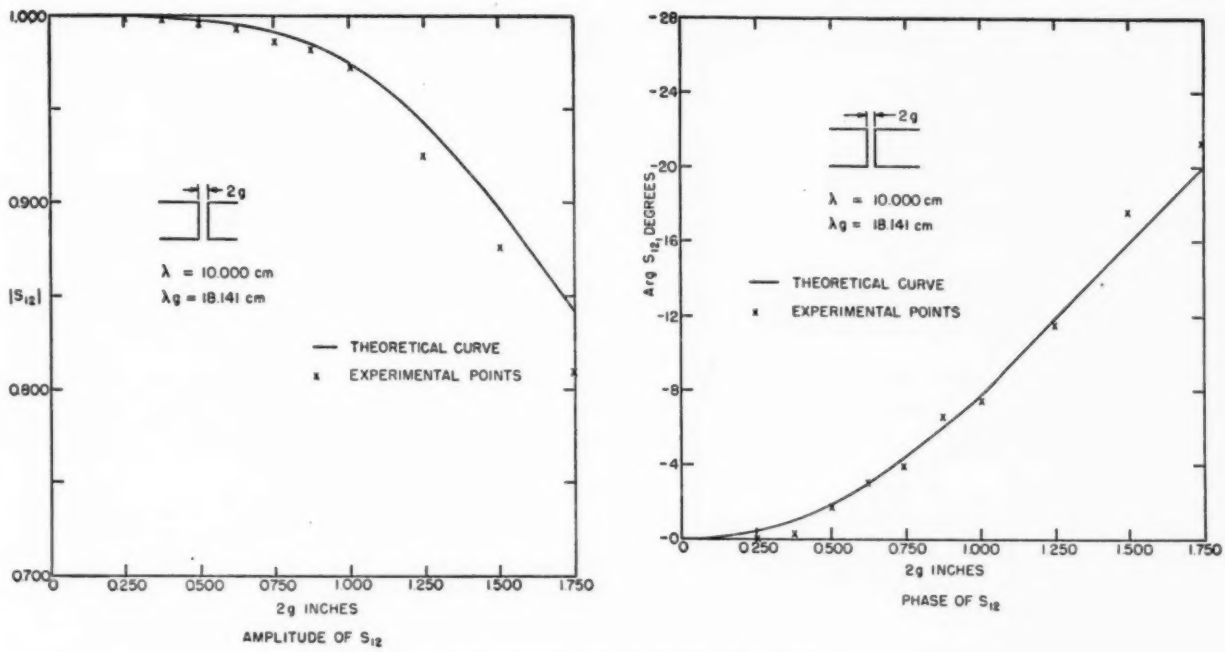


FIG. 6. Measured and calculated values of S_{12}.

Inserting (A5) into (A4) results in

$$\text{R.P.}\Delta_e \doteq \frac{\gamma_1 k^2 a^3}{\pi \rho_1^2} \int_0^1 \frac{(1-t^2)^{\frac{1}{2}}}{1+(\delta/1-t^2)} dt, \qquad \text{(A6)}$$

where the quantity δ is defined as

$$\delta = \frac{3}{8(ka)^2}.$$

Equation (A6) can be written as

$$\text{R.P.}\Delta_e \doteq \frac{\gamma_1 k^2 a^3}{4\rho_1^2}\left[1 - \frac{4\delta}{\pi}\int_0^1 \frac{(1-t^2)^{\frac{1}{2}}}{\delta+(1-t^2)} dt\right]. \qquad \text{(A7)}$$

Now since $\delta \ll 1$, it can be neglected in evaluating the remaining integral of (A7). One then obtains

$$\text{R.P.}\Delta_e \doteq \frac{\gamma_1 k^2 a^3}{4\rho_1^2}[1-2\delta] \doteq \frac{\gamma_1 k^2 a^3}{4\rho_1^2}. \qquad \text{(A8)}$$

To evaluate the imaginary part of Δ_e, examine the integral

$$\text{I.P.}\Delta_e = \text{I.P.}\frac{\gamma_1 a^2}{\pi^2 \rho_1^2} \times \int_{-\infty}^{\infty} \frac{[(2/\zeta g)J_1(\zeta g)]^2}{J_1[(k^2-\zeta^2)^{\frac{1}{2}}a]H_1^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a]} d\zeta. \qquad \text{(A9)}$$

Noting that the integrand is imaginary for values of $\zeta > k$, and that the major contribution to the integral occurs for $\zeta \gg k$, one can write

$$\text{I.P.}\Delta_e = \frac{\gamma_1 a^3}{\pi \rho_1^2} \int_{-\infty}^{\infty} \zeta \left[\frac{2}{\zeta g} J_1(\zeta g)\right]^2 d\zeta, \qquad \text{(A10)}$$

where the asymptotic expression

$$J_1[(k^2-\zeta^2)^{\frac{1}{2}}a]H_1^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a] \simeq J_1(i\zeta a)H_1^{(1)}(i\zeta a) \simeq \frac{1}{\pi i \zeta a} \qquad \text{(A11)}$$

with $|\zeta a| \gg 1$

has been used. It then follows that‡

$$\text{I.P.}\Delta_e \doteq \frac{8\gamma_1 a}{\pi \rho_1^2}\left(\frac{a}{g}\right)^2 \int_0^{\infty} \frac{[J_1(\zeta g)]^2}{\zeta} d\zeta = \frac{8\gamma_1 a}{\pi \rho_1^2}\left(\frac{a}{g}\right)^2 \int_0^{\infty} \frac{[J_1(t)]^2}{t} dt \qquad \text{(A12)}$$
$$= \frac{4\gamma_1 a}{\pi \rho_1^2}\left(\frac{a}{g}\right)^2.$$

In a similar fashion the real part of Δ_0 is given by

$$\text{R.P.}\Delta_0 = \text{R.P.}\frac{\gamma_1 a^2}{\pi^2 \rho_1^2}\left(\frac{8}{\gamma_1 g}\right)^2 \times \int_{-\infty}^{\infty} \frac{\left[\frac{2}{(\zeta g)^2}J_1(\zeta g) - \frac{J_0(\zeta g)}{\zeta g}\right]^2}{J_1[(k^2-\zeta^2)^{\frac{1}{2}}a]H_1^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a]} d\zeta. \qquad \text{(A13)}$$

‡ Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge University Press, Cambridge, 1945), p. 397.

It is seen again that the integrand has a real part when $k<\zeta$. Again assuming $kg\ll 1$, and using the leading term in the expansion of the numerator, one obtains

$$\text{R.P.}\Delta_0=\frac{2a^2}{\pi^2\rho_1^2\gamma_1}\int_0^k\frac{\zeta^2}{|H_1^{(1)}[(k^2-\zeta^2)^{\frac{1}{2}}a]|^2}d\zeta. \quad \text{(A14)}$$

Again making the substitution $t=\zeta/k$,

$$\text{R.P.}\Delta_0=\frac{2(ka)^3}{\pi^2\rho_1^2\gamma_1 a}\int_0^1\frac{t^2dt}{|H_1^{(1)}[(1-t^2)^{\frac{1}{2}}ka]|^2}. \quad \text{(A15)}$$

Introducing the asymptotic expression $|H_1^{(1)}[(1-t^2)^{\frac{1}{2}}\times ka]|^2\simeq 2/\pi ka(1-t^2)^{\frac{1}{2}}$ this becomes

$$\begin{aligned}\text{R.P.}\Delta_0&=\frac{(ka)^4}{\pi\rho_1^2\gamma_1 a}\int_0^1(1-t^2)^{\frac{1}{2}}t^2dt \\ &=\frac{(ka)^4}{16\rho_1^2\gamma_1 a}\ \text{(Peirce No. 145).}\end{aligned} \quad \text{(A16)}$$

Finally, by using an argument similar to that which led from (A9) to (A10), the imaginary part of Δ_0 is seen to be of the form

$$\text{I.P.}\Delta_0=\frac{128}{\pi\rho_1^2\gamma_1 a}\left(\frac{a}{g}\right)^4 I, \quad \text{(A17)}$$

where

$$I=\int_0^\infty\left\{\frac{4}{\zeta^3g^2}J_1^2(\zeta g)-\frac{4}{\zeta^2 g}J_1(\zeta g)J_0(\zeta g)+\frac{J_0^2(\zeta g)}{\zeta}\right\}d\zeta. \quad \text{(A18)}$$

By the substitution $t=\zeta g$,

$$I=\int_0^\infty\left\{\frac{4}{t^3}J_1^2(t)-\frac{4}{t^2}J_1(t)J_0(t)+\frac{J_0^2(t)}{t}\right\}dt. \quad \text{(A19)}$$

This integral can be expressed as

$$\begin{aligned}I&=\int_0^\infty\left\{\frac{J_1^2(t)}{t}-\frac{d}{dt}\left[\frac{J_1^2(t)}{t^2}+\frac{J_0(t)J_1(t)}{t}\right]\right\}dt \\ &=\int_0^\infty\frac{J_1^2(t)}{t}dt-\int_0^\infty\left[\frac{J_1^2(t)}{t^2}+\frac{J_0(t)J_1(t)}{t}\right]=\tfrac{1}{4}.\end{aligned} \quad \text{(A20)}$$

Therefore

$$\text{I.P.}\Delta_0=\frac{32}{\pi\rho_1^2\gamma_1 a}\left(\frac{a}{g}\right)^4. \quad \text{(A21)}$$

Finally, the resultant expressions for the variational parameters are given by

$$\begin{aligned}\Delta_e&\doteq\frac{\gamma_1 a}{\rho_1^2}\left\{\frac{(ka)^2}{4}+i\frac{4}{\pi}\left(\frac{a}{g}\right)^2\right\} \\ \Delta_0&\doteq\frac{1}{\gamma_1 a\rho_1^2}\left\{\frac{(ka)^4}{16}+i\frac{32}{\pi}\left(\frac{a}{g}\right)^4\right\}.\end{aligned} \quad \text{(A22)}$$

APPENDIX B. EVALUATION OF THE EXPLICIT ASYMPTOTIC EXPRESSION OF THE GREEN'S FUNCTION OF EQ. (11)

It is desired to derive an explicit form for the Green's function of Eq. (11) in order to evaluate the far-zone electric field in region I. The expression for the Green's function is given by

$$-\left[\frac{1}{r'}\frac{\partial}{\partial r'}r'G^{\text{I}}(r,r',z-z')\right]_{r'=a}=\frac{1}{2\pi}\int_C e^{+i\zeta(z-z')}\frac{J_1[(k^2-\zeta^2)^{\frac{1}{2}}r]}{J_1[(k^2-\zeta^2)^{\frac{1}{2}}a]}d, \quad \text{(B1)}$$

where the contour C is chosen as in Fig. 3.

It is noted that the integrand is analytic except for simple poles at $\zeta=\pm\gamma_1$ and $\zeta=\pm i\gamma_n(n>1)$. Since the result is required for $z\ll z'$ it is appropriate to close the contour with a large semicircle in the lower half-plane. It is apparent that the integral along the semicircle vanishes as its radius tends to infinity. Therefore, by the theory of residues, and noting that only the poles $\zeta=-\gamma_1$ and $\zeta=-i\gamma_n(n>1)$ are enclosed, one obtains

$$-\left[\frac{1}{r'}\frac{\partial}{\partial r'}r'G^{\text{I}}(r,r',z-z')\right]_{r'=a}=\frac{2\pi i}{2\pi}\sum_{\substack{\text{Residue at }\zeta=-\gamma_1\\ \zeta=-i\gamma_n\quad(n>1)}}. \quad \text{(B2)}$$

Since $e^{i\zeta(z-z')}$ represents a rapidly attenuated function when $z\ll z'$ and ζ is negative imaginary, all higher-order terms are negligible compared to the dominant-mode term.

Evaluating the residue corresponding to $\zeta=-\gamma_1$, one obtains the desired expression

$$-\left[\frac{1}{r'}\frac{\partial}{\partial r'}r'G^{\text{I}}(r,r',z-z')\right]_{r'=a}=\frac{i\rho_1 J_1\left(\rho_1\frac{r}{a}\right)}{\gamma_1 a^2 J_0(\rho_1)}\times e^{-i\gamma_1(z-z')}\quad z\ll z'. \quad \text{(B3)}$$

Impedance of a Top-Loaded Antenna of Arbitrary Length over a Circular Grounded Screen*

JAMES R. WAIT, *Radio Physics Laboratory, Defence Research Board, Ottawa, Canada,*

AND

WALTER J. SURTEES, *University of Toronto, Toronto, Canada*

(Received September 22, 1953)

The problem of a vertical monopole situated over a circular perfectly conducting screen lying on a finitely conducting ground is considered. An approximate method employed originally by Abbott to calculate the self-impedance is discussed. Using this formula explicit expressions are derived for the self-impedance of a thin, top-loaded vertical monopole. Sinusoidal current distribution is assumed.

INTRODUCTION

USUALLY the most important factor contributing to the inefficiency of low-frequency antennas is to be found in the ground system. The rules for ground system design are usually empirical and based on results of experiments on existing installations. Recently, the general problem has been investigated analytically by Abbott[1] who has developed a design procedure to select the optimum number of radial conductors specified given the values of the electrical constants of the ground.

An important related problem is the actual change of input impedance of the antenna for different sizes of ground systems. This analysis has been carried out for a vertical antenna situated centrally over a perfectly conducting disk by Leitner and Spence[2] and more recently by Storer.[3] They, however, only considered the case when the surrounding medium was free space. A more appropriate situation is when the disk is lying on the surface of a homogeneous conducting half-space which corresponds to the ground. It is not surprising that the solution of this problem is in general very difficult.

FORMULATION OF PROBLEM

With reference to a cylindrical polar coordinate system (ρ,ϕ,z) the antenna of height h is coincident with the positive z axis. The circular screen of radius a lies in the plane $z=0$ which is also the surface of the ground. The conductivity and dielectric constant of the ground are denoted by σ and ϵ, respectively, and the dielectric constant of the air by ϵ_0. The permeability of the whole space is μ which is taken to be the same as free space.

Due to the obvious polar symmetry in this problem, Maxwell's equations take the following form for time dependence according to $\exp(i\omega t)$:

$$\rho E_z=\frac{\eta}{\gamma}\frac{\partial}{\partial\rho}(\rho H_\phi) \tag{1}$$

$$E_\rho=-\frac{\eta}{\gamma}\frac{\partial H_\phi}{\partial z} \tag{2}$$

$$\frac{\partial E_z}{\partial\rho}-\frac{\partial E_\rho}{\partial z}=i\mu\omega H_\phi, \tag{3}$$

where γ and η are the intrinsic propagation constant and characteristic impedance of the medium, respectively, and defined by

$$\gamma=[i\mu\omega(\sigma+i\omega\epsilon)]^{\frac{1}{2}}$$

and

$$\eta=[i\mu\omega/(\sigma+i\omega\epsilon)]^{\frac{1}{2}}.$$

A subscript zero is affixed to these quantities in order that they should pertain to the air. That is,

$$\gamma_0=i(\epsilon_0\mu)^{\frac{1}{2}}\omega=i2\pi/\lambda,$$

where λ is the free space wavelength in meters and

$$\eta_0=(\mu/\epsilon_0)^{\frac{1}{2}}\simeq 377 \text{ ohms}.$$

Following the method of Storer,[3] the difference between the impedance Z of the antenna over a finite screen and the impedance Z_0 for an infinite screen is denoted by ΔZ and is given by

$$\Delta Z=Z-Z_0=-\frac{\eta_0}{\gamma_0^2 I_0^2}\int_0^h\left(\frac{1}{\rho}\frac{\partial}{\partial\rho}\rho H_\phi^s\right)I(z)dz. \tag{4a}$$

In this equation H_ϕ^s is the change of the magnetic field H_ϕ for a finite screen from the magnetic field H_ϕ^∞ for an infinite screen. The current on the antenna is $I(z)$ and the base current is I_0. It then follows that

$$\Delta Z=-\frac{1}{I_0^2}\int_0^\infty H_\phi^\infty(\rho,0)E_\rho(\rho,0)2\pi\rho d\rho, \tag{4b}$$

where

$$H_\phi^\infty(\rho,0)=-\frac{1}{2\pi}\frac{\partial}{\partial\rho}\int_0^h\frac{\exp[-\gamma_0(z^2+\rho^2)^{\frac{1}{2}}]}{(z^2+\rho^2)^{\frac{1}{2}}}I(z)dz. \tag{4c}$$

THE APPROXIMATE SOLUTION

The magnetic field in the ground outside the screen is a solution of the wave equation

$$(\Delta-\gamma^2)H_\phi(\rho,z)=0, \tag{5}$$

* Paper presented at Joint International Scientific Radio Union–Institute of Radio Engineers meeting in Ottawa, October 6th, 1953.

[1] F. R. Abbott, Proc. Inst. Radio Engrs. **40**, 846 (1952).

[2] A. Leitner and R. D. Spence, J. Appl. Phys. **21**, 1001 (1950).

[3] J. E. Storer, J. Appl. Phys. **22**, 1058 (1951).

and therefore

$$H_\phi(\rho,z)=\int_0^\infty J_1(\lambda\rho)e^{uz}p(\lambda)\lambda d\lambda, \quad z\leq 0, \tag{6}$$

where $u=(\lambda^2+\gamma^2)^{\frac{1}{2}}$ and the electric field is given by

$$E_\rho=-\frac{\eta}{\gamma}\int_0^\infty uJ_1(\lambda\rho)e^{uz}p(\lambda)\lambda d\lambda, \quad z\leq 0. \tag{7}$$

The binomial expansion of u is of the form

$$u=\gamma\left(1+\frac{\lambda^2}{2\gamma^2}+\cdots\right),$$

so that

$$E_\rho=-\frac{\eta}{\gamma}\left\{\int_0^\infty \gamma J_1(\lambda\rho)e^{uz}p(\lambda)\lambda d\lambda + \int_0^\infty (1/2\gamma)J_1(\lambda\rho)e^{uz}p(\lambda)\lambda^3 d\lambda+\cdots\right\}. \tag{8}$$

Now the differential equation for J is given by

$$\left(\frac{\partial}{\partial\rho}\frac{1}{\rho}\frac{\partial}{\partial\rho}\rho-\lambda^2\right)J_1(\lambda\rho)=0$$

and so it readily follows that, for $z=0$

$$E_\rho=-\eta H_\phi-\frac{\eta}{2\gamma^2}\frac{\partial}{\partial\rho}\frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho H_\phi) \pm\text{terms containing } \gamma^{-4}. \tag{9}$$

The second term is negligible if the propagation constant in the ground is sufficiently large and if ρH_ϕ is not varying too rapidly with ρ. That is ρH_ϕ should not change appreciably in a distance equal to $|\gamma^{-1}|$ in the radial direction for points outside the screen. This approximation is always well justified under practical conditions at low radio-frequencies.

Employing this result for E_ρ the equation for the self-impedance is now given by

$$\Delta Z=\frac{\eta}{I_0^2}\int_{\rho=a}^\infty H_\phi^\infty(\rho,0)H_\phi(\rho,0)2\pi\rho d\rho. \tag{10}$$

To apply this formula to an actual situation it can be assumed that $H_\phi(\rho,0)$ is not very different from $H_\phi^\infty(\rho,0)$ in the region of the ground plane where the losses are significant, and therefore

$$\Delta Z\simeq\frac{\eta}{I_0^2}\int_{\rho=a}^\infty [H_\phi^\infty(\rho,0)]^2 2\pi\rho d\rho. \tag{11}$$

An equivalent statement of this approximation is that the electrical radial current density at the surface of the ground is not appreciably affected by the finite conductivity of the soil. The approximate expression for ΔZ in Eq. (11) was first given in this form by Abbott[4] who quoted it without proof but pointed out that it was a plausible result. Later Monteath[5] also arrived at this same formula from an extension of the compensation theorem well known in circuit theory.

A better approximation to Eq. (10) might be to set

$$H_\phi(\rho,0)=H_\phi^\infty(\rho,0)F(p), \tag{12}$$

where $F(p)$ is Sommerfeld's "surface wave attenuation factor" which is a function of the numerical distance p which for a good conducting ground ($\sigma\gg\epsilon\omega$) is given by

$$p=(\pi\rho/\lambda)(\epsilon_0\omega/\sigma).$$

When $p\ll 1$ it can be shown or seen from tabulated numerical values[6] that $F(p)$ is very close to unity. If $p>1$ the magnitude of $F(p)$ drops appreciably below unity. However, since H_ϕ^∞ varies essentially as $e^{-\gamma_0\rho}/\rho$ and since $F(p)$ is a slowly varying function the contribution from the integrand in Eq. (11) for the cases where $p>1$ are negligible. At low radio-frequencies (<1000 kc) and for moderate ground conductivities ($\sigma>10^{-3}$ mhos per meter) the approximate formula given by Eq. (11) is sufficiently close to the formula in Eq. (10) to justify its use.

Up to this point no restriction has been placed on the current distribution on the antenna. For thin antennas it can usually be assumed that

$$I(z)=I_m\sin(\alpha-\beta z) \tag{13}$$

for

$$0\geq z\geq h,$$

where $\beta=\gamma_0/i=2\pi/\lambda$. The maximum current amplitude is I_m, and the conditions at the upper end of the antenna specify α. For unterminated ends $I(h)=0$ so that $\alpha=\beta h$ and then

$$I(z)=I_0\sin\beta(h-z)/\sin\beta h.$$

In some instances at low frequencies it is customary to load the antenna near the top so that effectively $I(h)\neq 0$. It is convenient then to set

$$\alpha=\beta(h+h'),$$

where h' specifies the degree of top loading.

Employing the value of $I(z)$ given by Eq. (13) and inserting it into Eq. (4c) the integration with respect to h can be carried out to give

$$H_\phi^\infty(\rho,0)=-\frac{iI_m}{2\pi}\left[\frac{e^{-i\beta r}}{\rho}\cos(\beta h-\alpha)-\frac{e^{-i\beta\rho}}{\rho}\cos\alpha - \frac{ih}{\rho}\frac{e^{-i\beta r}}{r}\sin(\beta h-\alpha)\right], \tag{14}$$

[4] F. R. Abbott, Standard Radial Ground Systems for l.f. and m.f. Monopole Transmitting Antennas. U. S. Navy Electronics Laboratory (San Diego), Report No. 219, Oct. 1950.

[5] G. D. Monteath, Proc. Inst. Elect. Engrs. (London) **98**, Pt. IV, 23 (1951).

[6] K. A. Norton, Proc. Inst. Radio Engrs. **24**, 1367 (1936).

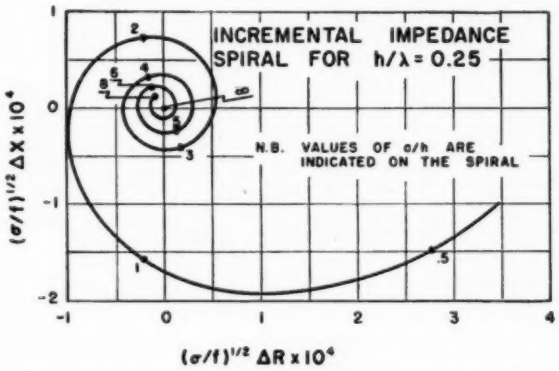


FIG. 1. The incremental self-impedance ΔZ for a quarter-wave monopole over a circular grounded screen.

where $r=(\rho^2+h^2)^{\frac{1}{2}}$. Using this value for H_ϕ^∞ the integration indicated in Eq. (11) can now be carried out to yield

$$\Delta Z=-\frac{\eta}{2\pi\sin^2\alpha}\{\cos^2(\beta h-\alpha)I_5+\cos^2\alpha I_6 -h^2\sin^2(\beta h-\alpha)I_1+2ih\cos\alpha\sin(\beta h-\alpha)I_3 -2\cos(\beta h-\alpha)\cos\alpha I_4-2ih\sin(\beta h-\alpha) \times\cos(\beta h-\alpha)I_2\}, \quad (15)$$

where I_1, I_2, $I_3\cdots I_6$ are integrals which can be expressed in terms of the exponential integral

$$Ei(-um)=-\int_u^\infty\frac{e^{-mx}}{x}dx$$

as follows:

$$I_1=\int_{\rho=a}^\infty\frac{e^{-2i\beta r}}{r^2\rho}d\rho=\frac{1}{h^2}E_i(-2i\beta r_0) -\frac{1}{2h^2}\{e^{-2i\beta h}Ei[-2i\beta(r_0-h)] +e^{2i\beta h}Ei[-2i\beta(r_0+h)]\}$$

$$I_2=\int_{\rho=a}^\infty\frac{e^{-2i\beta r}}{r\rho}d\rho=\frac{1}{2h}\{Ei[-2i\beta(r_0+h)]e^{2i\beta h} -Ei[-2i\beta(r_0-h)]e^{-2i\beta h}\}$$

$$I_3=\int_{\rho=a}^\infty\frac{e^{-i\beta(\rho+r)}}{r\rho}d\rho=\frac{1}{h}\{e^{i\beta h}Ei[-i\beta(g+1)h] -e^{-i\beta h}Ei[-i\beta(g-1)h]\}$$

$$I_4=\int_{\rho=a}^\infty\frac{e^{-i\beta(\rho+r)}}{\rho}d\rho=Ei(-i\beta hg) -Ei[-i\beta(g+1)h]e^{i\beta h}-Ei[-i\beta(g-1)h]e^{-i\beta h}$$

$$I_5=\int_{\rho=a}^\infty\frac{e^{-i2\beta r}}{\rho}d\rho=-\tfrac{1}{2}\{Ei[-2i\beta(r_0+h)]e^{i2\beta h} +Ei[-2i\beta(r_0-h)]e^{-i2\beta h}\}$$

$$I_6=\int_{\rho=a}^\infty\frac{e^{-2i\beta\rho}}{\rho}d\rho=-Ei(-2i\beta a),$$

where

$$r_0=(a^2+h^2)^{\frac{1}{2}} \quad\text{and}\quad g=[a+(a^2+h^2)^{\frac{1}{2}}]\frac{1}{h}.$$

The exponential integrals occurring in the above expressions are all of imaginary argument so therefore they can be expressed in terms of the sine and cosine integrals, $Si(x)$ and $Ci(x)$, respectively, as follows

$$Ei(-ix)=Ci(x)+i\left[\frac{\pi}{2}-Si(x)\right].$$

When the antenna is not terminated such that $h'=0$ or $\alpha=\beta h$ the expression for ΔZ simplifies to

$$\Delta Z=\frac{\eta}{4\pi\sin^2\beta h}\Big[Ei[-2i\beta(r_0+h)]e^{i2\beta h} +Ei[-2i\beta(r_0-h)]e^{-2i\beta h}+2\cos^2\beta hEi(-2i\beta a) +4\cos\beta h\{Ei(-i\beta hg)-Ei[-i\beta h(g+1)]e^{i\beta h} -Ei[-i\beta h(g-1)]e^{-i\beta h}\}\Big]. \quad (16)$$

It has been found that the results for ΔZ can be put in a very convenient form by writing

$$\Delta Z=\Delta R+i\Delta X$$

and then plotting ΔR as a function of ΔX for values of the ratio a/h. The curves take the form of spirals which are somewhat similar to the "impedance spirals" of Storer[3] for the antenna and disk situated in free space.

As a specific example the impedance spiral for a quarter wave vertical monopole (i.e. $h=\lambda/4$) without top loading is shown in Fig. 1. Since the displacement currents in the ground can be neglected at low radio-frequencies (i.e. $\omega\ll\sigma/\epsilon$) the curves can be conveniently normalized by multiplying ΔR and ΔX by $(\sigma/f)^{\frac{1}{2}}$. It is interesting to note that ΔR approaches infinity as a/h approaches zero. This is not surprising since the ground loss would be infinite if the lower end of the antenna were in direct contact with a homogeneous semi-infinite conductor.

Further calculations from Eqs. (15) and (16) have been carried out in this laboratory by Mr. D. A. Trumpler. In addition the effect of the ground screen on the radiation field has also been studied. These results are the subject of subsequent papers.[6,7]

ACKNOWLEDGMENT

The authors would like to thank Mr. H. Page, Mr. G. D. Monteath, and Dr. F. R. Abbott for their interesting and helpful comments in connection with this work.

[6] W. J. Surtees, Defence Research Board Report on Grant No. 67, Department of Electrical Engineering, University of Toronto (October, 1952).

[7] J. R. Wait and W. A. Pope, Paper No. 42.2, Institute of Radio Engineers Convention Record (March, 1954).

Plastic Deformation in Beams under Distributed Dynamic Loads*

J. A. SEILER AND P. S. SYMONDS
Brown University, Providence, Rhode Island
(Received June 25, 1953)

Plastic deformations in beams caused by symmetric dynamic loads distributed over finite lengths are considered in this paper. The analysis is based on the assumption of "plastic-rigid" behavior, so that the results should be valid when plastic deformations are sufficiently large. The final deformations for distributed loads are found not to differ much qualitatively from those to be expected for concentrated loads. However, the magnitude of the deformation decreases rapidly as the length of the loaded area increases from zero, for given load pulses. Thus the simplifying assumption of a mathematically concentrated load may cause a considerable overestimate of the final deformation.

1. INTRODUCTION

IN a recent paper[1] a method was described for calculating the permanent deformation of a bar of ductile metal subjected to a transverse impact load concentrated at the midpoint of the bar. The basis of the method, originally suggested by Prager, is that elastic strains are neglected entirely, the bar being assumed to be rigid except at cross sections where the limit, or fully plastic, moment M_0 is maintained. At such sections it is assumed that plastic strains of indefinitely large magnitude can occur. Since discontinuities in slope angle can take place at such sections, while the bending moment remains essentially constant there, these sections are conveniently termed "plastic hinges." This concept has been found to have valuable applications in predicting static loads which produce failure of the "plastic collapse" type in civil engineering structures. Papers by Baker[2] and Symonds and Neal[3] may be consulted for descriptions of these applications, and for details concerning the significance and calculation of the limit moment magnitude.

Whether or not the concept of localized deformation at plastic hinge sections can be usefully applied in problems of dynamic loading depends largely on how much the energy actually absorbed during plastic deformation under a given dynamic load exceeds the maximum elastic strain energy which could be stored in the structure. A criterion based on this hypothesis was presented in reference 1. Experimental investigations designed to test the predictions of this so-called "plastic-rigid" type of dynamic analysis will be highly desirable, and it is hoped that the results of the present paper will be helpful in planning and interpreting such experimental work.

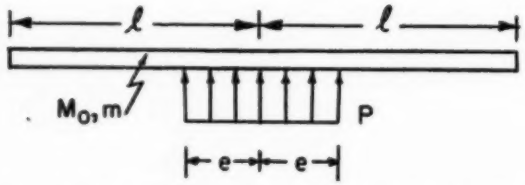


FIG. 1. Type of load distribution assumed in this paper.

It will be particularly desirable, in connection with experiments and with practical applications in general, to know how certain load characteristics may be expected to influence the results. In a previous paper,[4] the influence of characteristics of force pulses was investigated, the force being supposed mathematically concentrated at the midsection of the beam; as in reference 1, the case of a uniform beam with free ends was considered. In reference 4 a simple empirical formula was suggested by means of which the final central angle of deformation produced by a concentrated force could be estimated from knowledge of the total impulse and peak force value; the shape of the force-time curve was found to be relatively unimportant.

The present paper is concerned with the differences from the results of reference 4 to be expected when the load is distributed over a finite length of the beam rather than concentrated at the midsection. In order to obtain both qualitative and quantitative comparisons as simply as possible, the only form of load distribution curve used is that indicated in Fig. 1. The total load P is uniformly distributed over a length $2e$ where $e \leq l$. Thus the parameter specifying the distribution of the load is the ratio $e/l \equiv k$.

A previous and first study of this problem has been given by Salvadori and DiMaggio.[5] They used a smoother but more complicated load distribution function, the load intensity being written as

$$p(x)=\frac{P}{2l[1-e^{-c}(1+c/2)]}\left[\frac{c}{2}\left(1+c\frac{x}{l}\right)e^{-c(x/l)}\right], \quad (1)$$

where P is the total load, x is distance from midsection, and c is a concentration parameter which may vary from zero (corresponding to a uniform distribution over the entire beam) to infinity (corresponding to concen-

* The results presented in this paper were obtained in the course of research sponsored by the U. S. Office of Naval Research under Contract N7onr-35810 with Brown University.

[1] E. H. Lee and P. S. Symonds, J. Appl. Mech. **19**, No. 3, 308 (1952).

[2] J. F. Baker, J. Inst. Civil Engrs. (London) **27**, 188 (1949).

[3] P. S. Symonds and B. G. Neal, J. Franklin Inst. **252**, No. 5 383–407 (1951); No. 6, 469–492 (1951).

[4] P. S. Symonds, J. Appl. Mech. **20**, 475–481 (1953).

[5] M. G. Salvadori and F. DiMaggio, Quart. Appl. Math. **11**, 223–230, (1953).

tration of the load at the midsection). The authors carried their study only up to the point of showing that a sufficiently "spread out" load, characterized by a value of the parameter c less than 3.45, would introduce a qualitative difference in the behavior of the beam. This apparent difference will be discussed later in the present paper. With a simpler form of load distribution, as used in this paper, it is not difficult to carry the investigation further, and to obtain quantitative comparisons between plastic deformations occurring with various degrees of localization. These quantitative comparisons and more complete qualitative comparisons between final plastic deformations are the main purpose of the present paper. The main practical result is that a small degree of spreading of the load over a finite length of the beam leads to a considerable reduction of the total deformation at the midsection. Qualitatively, however, the present analysis shows that the *final shape* of the deformed beam will not differ radically from that predicted by analyses assuming the force to be concentrated at the midsection.

2. ANALYSIS OF EARLIER STAGES OF DEFORMATION

Taking the load to be uniformly distributed over the length $2e$ as shown in Fig. 1, we consider the successive phases of motion and deformation which can occur. We suppose the total load P increases steadily up to a maximum value and then decreases to zero, so that force pulses of the general shape shown in Fig. 2 are considered.

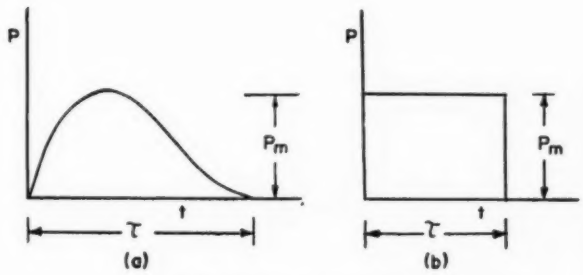


FIG. 2. Typical force pulse loadings.

Under these conditions, as the force increases from zero, its first effect is to cause a translation of the bar parallel to itself. Then all points of the bar have the acceleration $a=P/2ml$, where m is the mass per unit length. This initial translation is termed the "first phase" of the motion. The constant acceleration has the effect of a uniformly distributed loading, and the bending moment at any section x [positive bending moments defined as in Fig. 3(b)] is given by

$$\begin{aligned} x>e: \quad & M(x)=-\frac{m(l-x)^2}{2}a, \\ x<e: \quad & M(x)=-\frac{m(l-x)^2}{2}a+\frac{P}{2e}\frac{(e-x)^2}{2}, \end{aligned} \tag{2}$$

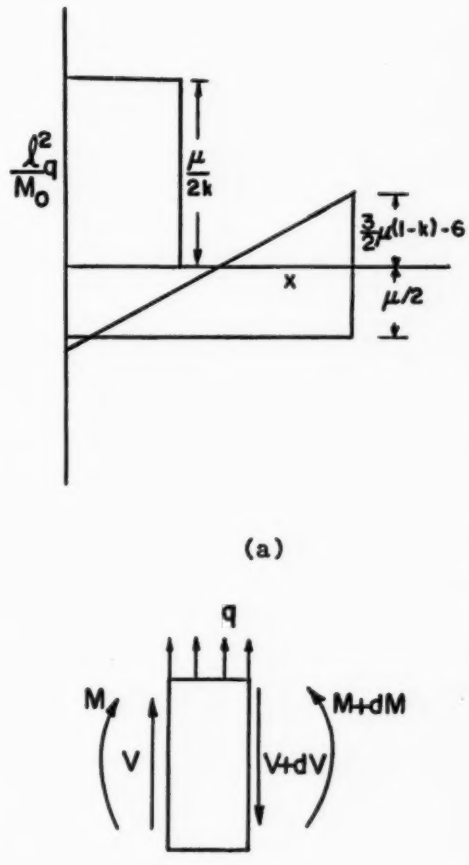


FIG. 3. Effective load components in the second phase.

where x is measured from the middle of the beam. The bending moment in this phase has a minimum at $x=0$ given by

$$M(0)=-\frac{ml^2a}{2}+\frac{Pe}{4}=-\frac{Pl}{4}+\frac{Pe}{4}. \tag{2a}$$

This first phase ends when the central moment reaches the value $-M_0$, corresponding to a load P_I given by

$$\frac{P_I l}{M_0}=\frac{4}{1-e/l}; \quad \text{or} \quad \mu_I=\frac{4}{1-k}, \tag{3}$$

where, for convenience, we define $Pl/M_0=\mu$ and $e/l=k$.

In the "second phase" of the motion, the two halves remain rigid but rotate with respect to each other, as in Fig. 4(a). Let a_0 denote the linear acceleration of the center, and $\alpha_0=d^2\theta_0/dt^2$ denote the angular acceleration of the right-hand half (positive clockwise). Assuming θ_0 to remain small, the equations of motion are

$$\frac{P}{2}=ml(a_0-\tfrac{1}{2}l\alpha_0); \quad \frac{P}{2}\left(\frac{l}{2}-\frac{e}{2}\right)-M_0=\frac{ml^2\alpha_0}{12}. \tag{4}$$

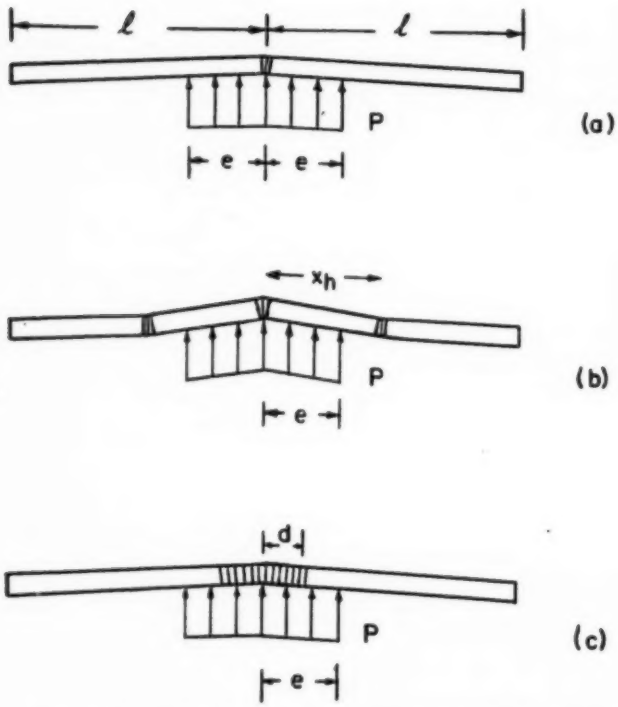


FIG. 4. Possible types of deformation: (a) second phase; plastic hinge at mid-point only; (b) third phase with additional hinges at interior points in each half-beam; (c) third phase with development of *finite* central plastic region.

Solving these for a_0 and α_0, we obtain

$$a_0=\frac{M_0}{ml^2}[\mu(2-\tfrac{3}{2}k)-6]$$
$$l\alpha_0=\frac{M_0}{ml^2}[3\mu(1-k)-12]. \tag{5}$$

This second phase continues until the moment at some additional cross section reaches the magnitude M_0.

To see how this may occur, it is helpful to refer to typical diagrams of effective load, shear force, and bending moment in the right-hand half of the beam. The effective load intensity q, with sign defined as in Fig. 3(b), can be regarded as made up of three parts, as shown in Fig. 3(a); these are the intensity $P/2e=(M_0/l^2)(\mu/2k)$ of the given load distributed over the length e, the uniform load intensity $-ma(l/2)=-(M_0/l^2)(\mu/2)$ corresponding to the acceleration $a(l/2)$ of the center of gravity of the half-beam, and finally the linearly varying effective load intensity corresponding to the angular acceleration α_0 of the half-beam, whose maximum value is $ma_0(l/2)=M_0/l^2\times[\frac{3}{2}\mu(1-k)-6]$. q, V, M curves are shown in Fig. 5.

Two limiting conditions are evident from the diagrams of Fig. 5. In the first place, unless $q(l)$ is positive, the V curve cannot cut the axis in the unloaded position $x>e$, and hence there can be no positive moment in the beam. Hence, in order for it to be possible to develop a plastic hinge at a point where the maximum moment reaches the value $+M_0$, we must have μ and k related so that

$$\tfrac{3}{2}\mu(1-k)-6-\frac{\mu}{2}>0 \tag{6}$$

or

$$\mu(1-\tfrac{3}{2}k)>6. \tag{6a}$$

Secondly, at the left-hand end of the half-beam, $q(0)$ must remain positive; if it did not, V would have negative values in the neighborhood of $x=0$, and hence $M(x)$ would decrease below $-M_0$ in this region. This requires the following relation between μ and k:

$$\frac{\mu}{2k}-\tfrac{3}{2}\mu(1-k)+6-\frac{\mu}{2}>0 \tag{7}$$

or

$$\mu(3k-1)(1-k)<12k. \tag{7a}$$

It will then be evident that the second phase may be terminated in some cases because there is a positive moment in the interior of the half-beam [as in Fig. 5(c)] which reaches the value M_0 at a sufficiently large value of the load; and in other cases because the moment in the neighborhood of the midsection reaches the negative limit moment value $-M_0$ over a finite length of the beam, as in Fig. 4(c).

The first case, occurrence of a "positive" plastic hinge in the interior, is possible only if the inequality (6) is satisfied. From (6a), therefore, it is evident that in order for this new hinge to form in the interior of each half-beam it is necessary to have

$$k<\tfrac{2}{3}. \tag{8}$$

The second case, of spreading of the central hinge, can occur only if inequality (7) is *not* satisfied, and hence only if

$$k>\tfrac{1}{3}, \tag{9}$$

since inequality (7a) will be satisfied, regardless of μ, if $k\leq\frac{1}{3}$.

In the case of a concentrated force,[1,4] the second phase ends when the positive bending moment reaches the value M_0, and this always occurs when the load parameter μ reaches the value $\mu_{II}=22.89$. The considerations of the preceding paragraph show that when the load is sufficiently spread out ($k\geq\frac{2}{3}$), the formation of interior positive hinges will not take place in this manner. This is the main qualitative difference caused by loading over a finite length of the beam. Instead of new outboard hinges making their appearance, for sufficiently large k values the new effect is the spreading of the central hinge, which in the earlier part of the motion is always concentrated (according to the basic hypotheses of the method) at a single section.

We now put in more precise terms the conditions for (1) occurrence of a positive plastic hinge in the interior of the beam, and (2) spreading of the central negative plastic hinge.

(1) Occurence of lateral plastic hinge:
For $e \leq x \leq l$, the bending moment is given by

$$M(x) = \frac{P}{2}\left(x - \frac{e}{2}\right) - M_0 - \frac{mx^2\alpha_0}{2} + \frac{mx^3\alpha_0}{6}. \quad (10)$$

This moment will have a stationary maximum at a point which can be found by solving the equation

$$V = \frac{dM}{dx} = \frac{P}{2} - mxa_0 + \frac{mx^2\alpha_0}{2} = 0. \quad (11)$$

Using Eqs. (5) this yields

$$\frac{x}{l} = \frac{\mu}{3\mu(1-k) - 12}. \quad (12)$$

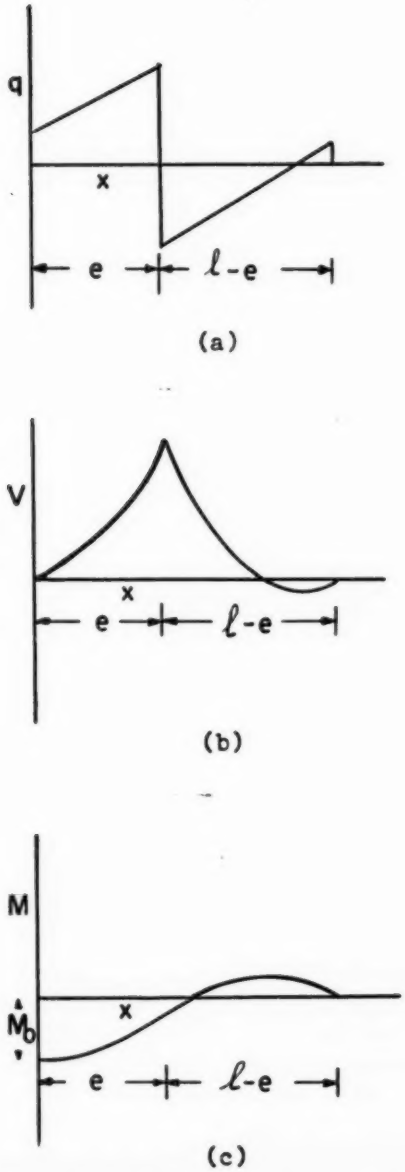


FIG. 5. Typical diagrams of effective load intensity and associated shear and bending moment diagrams in second phase.

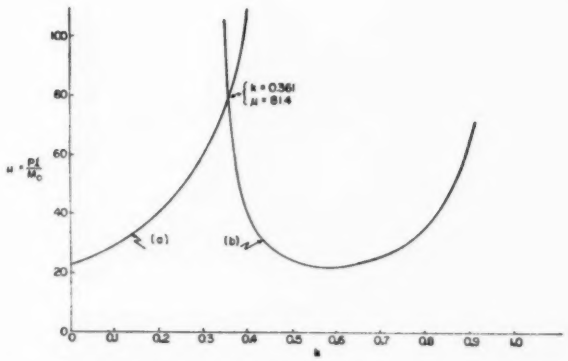


FIG. 6. Plots of μ at which the second phase ends as junction of $k = e/l$ corresponding to (a) appearance of lateral "positive" plastic hinges, and (b) spreading of the central hinge.

This value of x is put in Eq. (10) to give the expression for the maximum moment. Then the value of μ for which the second phase ends, is found by setting the resulting expression equal to M_0. The following cubic equation in μ_{II} and k is obtained

$$(3k-2)^2\mu_{II}^3 + 36(12k^2 - 18k + 7)\mu_{II}^2 + 423(5k-4)\mu_{II} + 3456 = 0. \quad (13)$$

For any particular value of k, the real root of Eq. (13) then gives the value of μ for which the lateral hinges first appear, on the assumption that the kinematic picture of Fig. 4(b) is correct.

(2) The central hinge must begin to spread if

$$\left.\frac{d^2M}{dx^2}\right|_{x=0} = \left.\frac{dV}{dx}\right|_{x=0} = q(0) = 0. \quad (14)$$

The critical relation is, therefore, obtained by replacing the inequality in (7a) by an equality, which yields

$$\mu_{II} = \frac{12k}{(3k-1)(1-k)}. \quad (15)$$

The curves of μ as a function of k according to Eqs. (13) and (15) are plotted in Fig. 6. It is found that the positive branch of the curve from Eq. (15) crosses that from Eq. (13) at $k = 0.361$, $\mu_{II} = 81.4$. We conclude, therefore, as follows:

(1) If $k < 0.361$, the second phase ends with the formation of outboard hinges. The initial location of these hinges will be found from Eq. (12), and the value of the load parameter μ at which they appear is the real root of Eq. (13). Plastic regions appear in the manner indicated in Fig. 4(b).

(2) If $k > 0.361$, the termination of the second phase will be due to the spreading of the central plastic hinge into a finite region in which the moment has the constant value $-M_0$. The load parameter at which this effect begins is given by Eq. (14), and the corresponding type of deformation is indicated in Fig. 4(c).

This critical value $k = 0.361$ corresponds to the critical value $c = 3.45$ found by Salvadori and DiMaggio for the

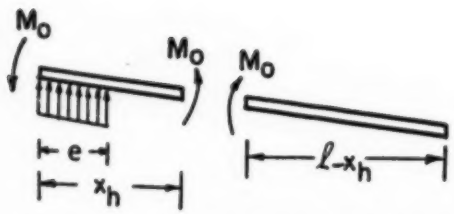


FIG. 7. Free body diagrams for motion in third phase with concentrated central hinge.

smooth load distribution used by them. They refer to "splitting of the central hinge" as the effect resulting when Eq. (14) is satisfied before the lateral hinges form, and they point out that the moment has the constant magnitude M_0 in the central portion of the beam. The use of the term "spreading" seems preferable to "splitting," which might be taken to imply the division of the original central hinge into two distinct hinges separated by a region moving as a rigid body. A true splitting of this sort could occur if the intensity of the distributed load has a relative minimum at the center of the beam, but not otherwise.

For the analysis of the motion after the close of the second phase we assume first that the load distribution parameter $k=e/l$ is less than 0.361 and then that k exceeds 0.361.

3. ANALYSIS FOR $k<0.361$

In this case the lateral hinges appear while the central hinge is still localized at the center section. The appropriate free body diagrams for the right-hand half are shown in Fig. 7, the distance of the lateral hinge from the center being denoted by x_h. It is assumed that $x_h>e$. The equations of motion, as in references 1 and 4, are

$$\frac{P}{2}=mx_h(a_0-\tfrac{1}{2}x_h\alpha_0), \tag{16a}$$

$$\frac{P}{2}\left(\frac{x_h}{2}-\frac{e}{2}\right)-2M_0=\frac{mx_h^3\alpha_0}{12}, \tag{16b}$$

$$0=m(l-x_h)[a_1+\tfrac{1}{2}(l-x_h)\alpha_1], \tag{17a}$$

$$M_0=\frac{1}{12}m(l-x_h)^3\alpha_1. \tag{17b}$$

The fifth equation relating the five unknowns a_0, a_1, α_0, α_1, and x_h is that which expresses the discontinuity of linear acceleration at the hinge location x_h, which is, in general, a function of time. This is (see references 1 and 4)

$$a_0-x_h\alpha_0-[a_1+\alpha_1(l-x_h)]=\frac{dx_h}{dt}(\omega_0-\omega_1), \tag{18}$$

where ω_0, ω_1 are the angular velocities of the segments to the left and to the right, respectively, of the hinge, i.e.,

$$\alpha_0=\frac{d\omega_0}{dt},\quad \alpha_1=\frac{d\omega_1}{dt}. \tag{19}$$

Equations (16)–(18) yield the following system of equations:

$$\alpha_0=\frac{d\omega_0}{dt}=\frac{3M_0}{ml^3\xi^3}[\mu(\xi-k)-8], \tag{20a}$$

$$\alpha_1=\frac{d\omega_1}{dt}=\frac{12M_0}{ml^3(1-\xi)^3}, \tag{20b}$$

$$a_0=\frac{M_0}{2ml^2\xi^2}[\mu(4\xi-3k)-24], \tag{20c}$$

$$a_1=-\frac{6M_0}{ml^2(1-\xi)^2}, \tag{20d}$$

$$\frac{M_0}{ml^3}\left[\frac{\mu}{\xi}-\frac{3}{2}\frac{\mu k}{\xi^2}-\frac{12}{\xi^2}+\frac{6}{(1-\xi)^2}\right]=-\frac{d\xi}{dt}(\omega_0-\omega_1), \tag{20e}$$

where

$$\xi=\frac{x_h}{l}. \tag{21}$$

These equations hold so long as the bending moment has a magnitude less than M_0 everywhere except at the sections $x=0$ and $x=x_h$. The effective load, shear, and moment diagrams for the segments to the left and right of the moving hinge, (Fig. 8) indicate that the critical condition will be a spreading of the central hinge. This will occur when the following condition is satisfied:

$$\left.\frac{d^2M}{dx^2}\right|_{x=0}=\left.\frac{dV}{dx}\right|_{x=0}=q(0)=0. \tag{22}$$

From Eq. (16) this leads to the condition

$$\mu=\frac{24k}{(\xi-k)(3k-\xi)}. \tag{23}$$

The use of the limiting condition requires that Eqs. (20) first be solved, yielding ξ (among other quantities), as a function of μ; these solutions will be valid as long as μ is less than the value computed from Eq. (23).

If any given finite force pulse is replaced by a rectangular pulse of the same peak force and the same area, i.e., total impulse, the final deformations computed fo the rectangular pulse will exceed the true ones for the original pulse, but for many purposes the results will be useful estimates (e.g., in error by 15 or 20 percent) (see reference 4). Therefore, to obtain quantitative comparisons of total deformations with various values of k, we have assumed a rectangular force pulse shape, Fig.

2(b). Since the details of the method of analysis exactly parallel those in reference 4, only the final results are shown in Fig. 9, where curves of the deformation parameters $ml^2\theta_0/M_0T^2$ are plotted as functions of $\mu_m \equiv P_m l/M_0$ for a series of values of the distribution parameter k. These results and others are discussed further in the final section of the paper.

4. ANALYSIS FOR $k > 0.361$

In this case the central hinge begins to spread before lateral hinges appear. At time t the central hinge will have spread out a distance d from the center; in the region $-d \leq x \leq d$ the moment is constant at the value $-M_0$. The equations of motion for the segment of length $l-d$ are obtained from the free body diagram of

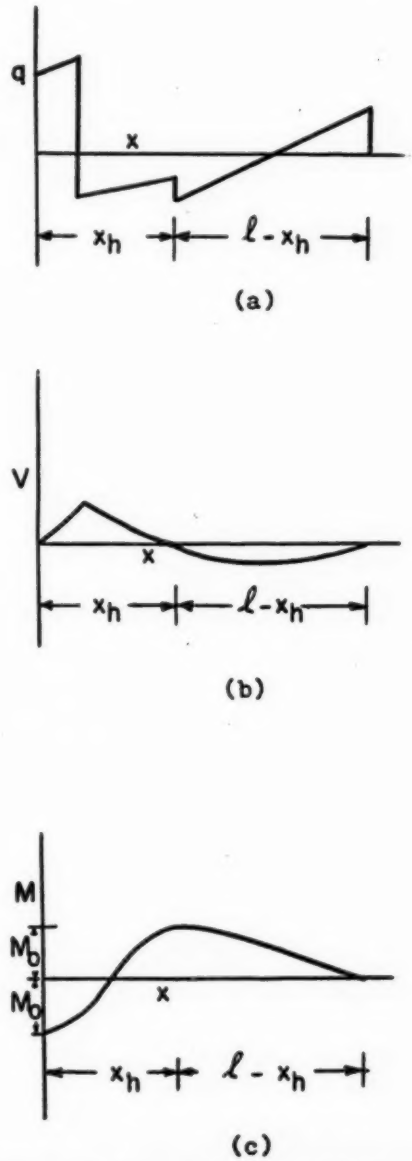


FIG. 8. Diagrams of effective load, shear, and bending moment in the third phase, with concentrated central hinge.

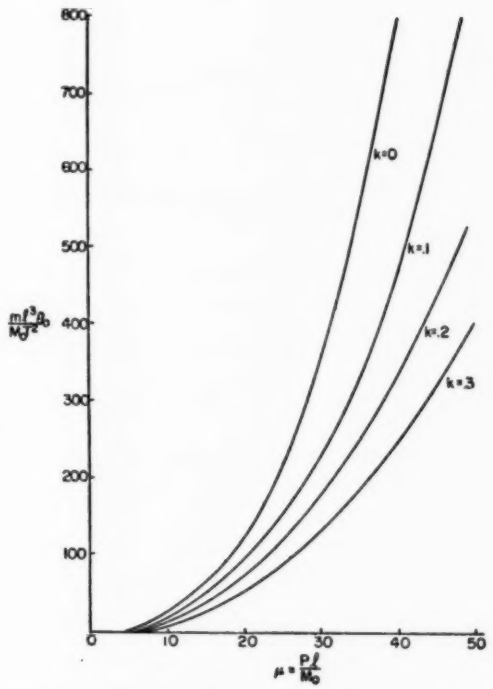


FIG. 9. Final central angle of deformation caused by various values of the load distribution parameter k, as function of maximum load P of a rectangular force pulse.

Fig. 10, and are found to be

$$\frac{P}{2e}(e-d) = m(l-d)[a_0' - \tfrac{1}{2}(l-d)\alpha_0'] \tag{24a}$$

$$\frac{P}{2e}(e-d)\left[\frac{l-d}{2} - \frac{e-d}{2}\right] - M_0 = \frac{1}{12}m(l-d)^3\alpha_0', \tag{24b}$$

where a_0' is the acceleration at $x=d$ and α_0' is the angular acceleration of the segment of length $l-d$.

Equations (24) yield expressions for the accelerations as follows:

$$a_0' = \frac{M_0}{ml^2(1-\lambda)^2}\left[\frac{\mu}{2k}(k-\lambda)(4-\lambda-3k) - 6\right], \tag{25a}$$

$$\alpha_0' = \frac{12M_0}{ml^2(1-\lambda)^3}\left[\frac{\mu}{4k}(k-\lambda)(1-k) - 1\right], \tag{25b}$$

where

$$\lambda \equiv \frac{d}{l}. \tag{25c}$$

To determine d, we use the fact that the central hinge spreads at a rate just sufficient to prevent the moment near $x=d$ from decreasing below the value $-M_0$. The condition is that

$$\left.\frac{d^2M}{dx^2}\right|_{x=d} = \left.\frac{dV}{dx}\right|_{x=d} = q(d) = \frac{P}{2e} - ma_0' = 0. \tag{26}$$

FIG. 10. Free body diagram of segment to right of central plastic region.

Using Eq. (25a) and simplifying, it is found that λ is the following function of μ and k:

$$\lambda=\frac{\mu(3k-1)(1-k)-12k}{2\mu(1-k)}. \tag{27}$$

Plots of μ *vs* λ for various values of k are given in Fig. (11). Thus for $k>0.361$ the critical portion of the beam ($-x\leq\lambda l\leq x$) has the constant moment $-M_0$, so that the accelerations of all elements are proportional to the local load intensity $P/2e$. The two segments of length $l-d=l(1-\lambda)$ rotate as if rigid bars hinged at the sections $x=\pm d$ [Fig. 4(c)]. So long as the moment in the outer segments remains less than M_0 in magnitude, this is the correct picture. The load magnitude at which this picture ceases to be the correct one is that at which new "positive" hinges form in the interior of the outer segments. The maximum bending moment in these segments can be investigated by the same method as was used in the case of a localized central hinge. The following equation was obtained by setting the maximum bending moment in one segment $d\leq x\leq l$ equal to M_0. It is analogous to Eq. (16) but depends now upon $\lambda=d/l$ in addition to k.

$$2=\frac{\mu^2(k-\lambda)^2(1-\lambda)^2[\mu(k-\lambda)(2+\lambda-3k)-12k]}{4k[3\mu(k-\lambda)(1-k)-12k]^2}$$
$$-\frac{\mu(k-\lambda)^2}{4k}+\frac{\mu^3(k-\lambda)^3(1-\lambda)^3}{6k[3\mu(k-\lambda)(1-k)-12k]^3}. \tag{28}$$

To determine the load at which the additional hinges first appear, the above equation must be solved simultaneously with Eq. (27). Alternatively, for any chosen value of k (>0.361), plots of μ *vs* λ from Eq. (28) and Eq. (27) can be drawn, and the intersection of the two curves yields the initial value of μ in question. As one example, the case $k=0.5$ has been worked out. In this case, when the two equations are combined we obtain

$$\mu^4-192\mu^3+1152\mu^2-110592=0. \tag{29}$$

This has the solution $\mu=186$, corresponding to $\lambda=0.242$; the new hinge appears at $x=0.660l$.

5. NUMERICAL RESULTS AND CONCLUSIONS

The type of load distribution considered here, with uniform load intensity over part of the beam and zero load over the rest, is intended less as a realistic loading than as the simplest type of distributed load to handle analytically. It seems intuitively clear that the qualitative results deduced from this simple case will apply to general cases of loads spread over finite areas. The main qualitative result is, in fact, that the general picture of the final deformation is exactly the same as that in the case of concentrated loading unless the length over which the load is applied exceeds about a third of the total length of the beam. Only when k exceeds the value 0.361 does a qualitative difference appear. This consists of a spreading of the central hinge, so that instead of a plastic hinge localized at the mid-section there is a finite length of the beam in which the bending moment has the magnitude M_0. However in real beams a hinge cannot be localized at a single cross section in any case since this would imply infinite strains; the plastic region is always spread over a finite region even when the load is strongly localized. Thus the final plastic deformation of a beam subjected to a sufficiently large distributed load will not in any case differ radically from that predicted by an analysis using a mathematically concentrated load.

As far as quantitative comparisons are concerned it is of particular interest to investigate how a small degree of load spreading modifies the magnitude of the final deformation predicted by the simpler analysis for a concentrated force, since in practice loads are always applied over finite areas.

To provide information of this kind calculations have been made of the final central angle θ_0 produced by a series of loads with k values ranging from zero to 0.30. The force-time curve was taken, for maximum simplicity, to have a square-wave shape, as in Fig. 2(b).

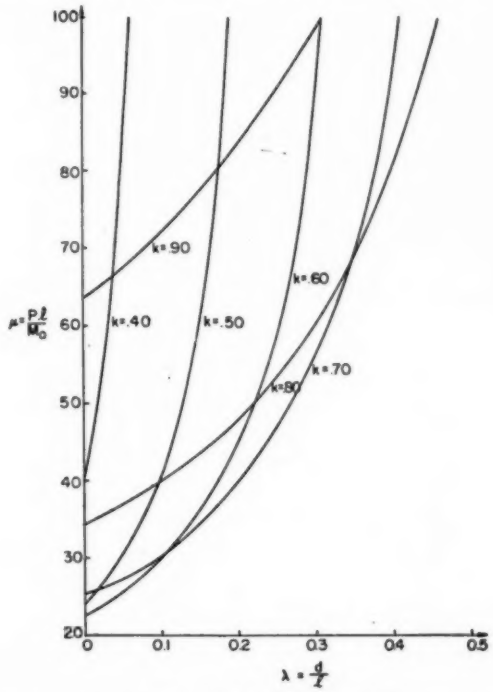


FIG. 11. Dependence of width d of central hinge on load P for various values of load distribution parameter $k=e/l$.

For this type of force pulse the lateral hinges move instantaneously to positions which they maintain so long as the load remains constant. The lateral hinge coordinate ξl is found as the solution of the following equation:

$$\frac{\mu}{\xi}-\frac{3\mu k}{2\xi^2}-\frac{12}{\xi^2}+\frac{6}{(1-\xi)^2}=0. \tag{30}$$

This is Eq. (20e) with the right-hand side set equal to zero. Figure 12 shows curves of ξ plotted against μ for values of k ranging from zero up to 0.30. The calculation of the further increments of deformation which occur after the load drops to zero are made in the manner described in reference 4.

Figure 9 shows curves of the dimensionless parameters $ml^2\theta_0/M_0T^2$ plotted as a function of the dimension-

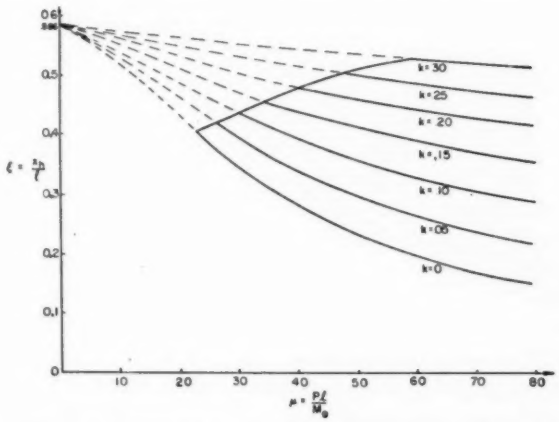


FIG. 12. Coordinate of lateral hinge x_A for various values of load distribution parameter k, for constant load P.

less load parameter $\mu_m=P_ml/M_0$, for k equal to 0, 0.10, 0.20, and 0.30. Figure 13 shows curves of $ml^2\theta_0/M_0T^2$ plotted against k for various values of μ_m. These curves show that when the maximum load is large (e.g., $\mu_m>40$) the deformation falls off very rapidly. For example, with $\mu_m=40$, the deformation parameter drops from about 800 for a concentrated load ($k=0$) to about

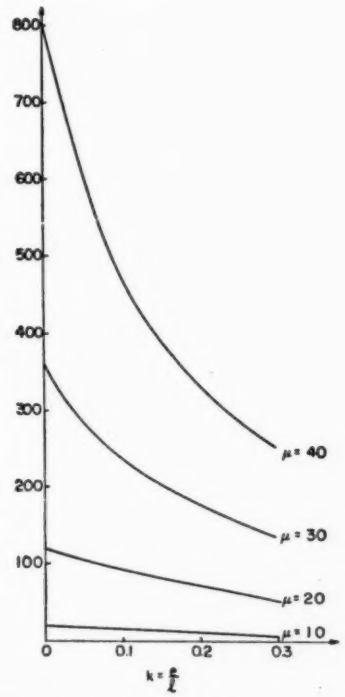


FIG. 13. Final central angle of deformation as function of load distribution parameter k.

600 for $k=0.05$. This apparently large reduction caused by the distribution of the load over a finite length may be somewhat exaggerated, as compared with practical cases, because of the special conditions of the present calculation, i.e., the assumption of a rectangular force pulse and of a constant load intensity over the loaded segment of the beam. However, the results indicate that the assumption of a mathematically concentrated load may lead to an appreciable overestimate of the deformation, since the actual load is spread over a small but finite length of the beam. This result should be of significance in evaluating experimental results. It should perhaps also be noted that strain-hardening will cause a further reduction in the permanent deformation; this effect will presumably be more important the higher the degree of concentration of the load.

Tube Noise under Large Transit-Time Conditions*

S. V. YADAVALLI
Electronics Research Laboratory, University of California, Berkeley, California
(Received July 8, 1953)

That the noise convection current in a diode would be the same as the noise current in a triode under the conditions $\theta_1 \gg 1$ and $\theta_2 \ll 1$ (θ_1 and θ_2 are transit angles in radians of the potential minimum-to-grid and grid-to-plate regions, respectively) is shown by following the approach indicated by Rack and applying the Llewellyn-Peterson small signal theory to study the noise in a triode. Although the general trend of the experimental results agrees very well with the prediction from the Rack approach, the measured value of noise power is found to be about 2 db below the predicted value. The experimental results are applicable to triodes as well under certain conditions.

I. INTRODUCTION

In recent years many workers[1–5] have studied the signal and noise behavior of high-frequency tubes on the assumption that the velocity of electrons is single valued thus neglecting the effect of the thermal velocity distribution. In general, analyses such as the above have been quite successful in predicting the observed results. The effect of transit time on shot noise in vacuum tubes is studied in the present investigation by the familiar Rack-Llewellyn-Peterson method.[1,6] The purpose of this paper is to investigate the behavior of noise convection current in a diode and electron streams under large transit time conditions. It may be remarked here, however, that the present results are applicable to triodes as well under certain conditions.

The Rack-Llewellyn-Peterson† method employs a set of equations for a diode region obtained by Llewellyn.[7] Llewellyn and Peterson[1] have shown that the ac behavior of a diode-type region can be described by specifying certain boundary conditions at one of the two planes bounding the diode region. The assumptions made in the above analysis are given elsewhere‡ and will not be repeated here. The R-L-P method essentially consists of the following: first, the spectral density of the mean-square fluctuations of the average velocity of the electrons at the potential minimum is obtained; then, instead of considering a multivelocity electron stream from the potential minimum on, a single velocity stream having the same mean-square fluctuations is considered. A procedure such as Rack's which is based upon a hypothesis must be supported by experimental results or more detailed analysis.

In recent years, experimental investigations have been carried out by Duvall,[8] Kompfner, *et al.*,[9] and Lavoo[10] on noise in diodes under fairly large transit time conditions, and by Cutler and Quate,[11] Watkins,[5] and others regarding noise in beam-type tubes.

II. ANALYSIS

Let us investigate a triode and find how the noise will vary with the cathode-to-grid transit angle. The region between the cathode and the potential minimum is neglected since otherwise this would contradict the assumption that the velocity of the electron stream is single valued. Only space-charge limited emission is considered below since this is of greatest interest in practice.

Consider a parallel-plane space-charge limited diode region whose entrance plane a is taken to be just beyond the potential minimum and whose exit plane g is infinitesimally close to the actual position of the grid wires. Considering a small frequency interval of width Δf, one can write the following electronics equations:§

$$\begin{aligned} I_g &= a_{11}(V_g - V_a) + a_{12}q_a + a_{13}v_a \\ q_g &= a_{21}(V_g - V_a) + a_{22}q_a + a_{23}v_a \\ v_g &= a_{31}(V_g - V_a) + a_{32}q_a + a_{33}v_a. \end{aligned} \tag{1}$$

The coefficients a_{11}, $a_{12}\cdots a_{33}$ are functions of frequency and space charge, but, if the frequency interval of the measuring apparatus is small, they can be considered to be constant throughout the range of frequencies between f and $f+\Delta f$. $(V_g - V_a)$ is the fluctuation in potential across the diode, I is the fluctuation in total current (convection plus displacement currents) passing through the diode region $a-g$, q_a, v_a and q_g, v_g are fluctuations in the convection current density and velocity at the planes a and g respectively (see Fig.

* Based on a portion of a dissertation submitted by the author to the University of California. Reference is made to this dissertation for complete details.

[1] F. B. Llewellyn and L. C. Peterson, Proc. Inst. Radio Engrs. **32**, 144 (1944).

[2] L. C. Peterson, Proc. Inst. Radio Engrs. **35**, 1264 (1947).

[3] J. R. Pierce, *Traveling Wave Tubes* (D. Van Nostrand Company, Inc., New York, 1950), Chap. X.

[4] L. D. Smullin, Tech. Report No. 142, Research Laboratory of Electronics, Massachusetts Institute of Technology (1949).

[5] D. A. Watkins, Proc. Inst. Radio Engrs. **40**, 65 (1952).

[6] A. J. Rack, Bell System Tech. J. **17**, 592 (1938).

† From now on we shall use the abbreviation R-L-P method for Rack-Llewellyn-Peterson method.

[7] F. B. Llewellyn, *Electron Inertia Effects* (Cambridge University Press, Cambridge, 1941).

‡ See, for instance, reference 2.

[8] G. E. Duvall, Tech. Report No. 82, Research Laboratory of Electronics, Massachusetts Institute of Technology (1948).

[9] K. Kompfner, *et al.*, J. Inst. Elec. Engrs. (London) **93**, Pt. 3A, No. 9 (March-May 1946).

[10] N. T. Lavoo, Proc. Inst. Radio Engrs. **37**, 383 (1949).

[11] C. C. Cutler and C. F. Quate, Phys. Rev. **80**, 875 (1950).

§ The Peterson form of the electronics equations is used here, see for instance reference 2.

1). Complete space charge in the cathode-to-grid region is assumed in the following treatment. According to the notation of Llewellyn and Peterson, the space-charge factor‖ $\xi_1=1$ for the complete space-charge limitation. Also the dc velocity u_a at the potential minimum is taken to be zero.

Using the above assumptions, one obtains for region (1) the following¶ from Eq. (1):

$$\begin{aligned} I_g &= a_{11}^{(1)}(V_g-V_a)+a_{13}^{(1)}(v_f)_a \\ q_g &= a_{21}^{(1)}(V_g-V_a)+a_{23}^{(1)}(v_f)_a \\ v_g &= a_{31}^{(1)}(V_g-V_a)+a_{33}^{(1)}(v_f)_a. \end{aligned} \tag{2}$$

Note from Eqs. (1) and (2), that if a small amount of initial fluctuation current is assumed to exist at the plane a the contribution from this fluctuation is negligible in comparison with that due to a fluctuation in velocity at the plane a. Now, treating again the grid-to-plate region as another diode we can write:

$$\begin{aligned} I_p &= a_{11}^{(2)}(V_p-V_g)+a_{12}^{(2)}[\alpha_1 q_g+N']+a_{13}^{(2)}v_g \\ q_p &= a_{21}^{(2)}(V_p-V_g)+a_{22}^{(2)}[\alpha_1 q_g+N']+a_{23}^{(2)}v_g \\ v_p &= a_{31}^{(2)}(V_p-V_g)+a_{32}^{(2)}[\alpha_1 q_g+N']+a_{33}^{(2)}v_g \end{aligned} \tag{3}$$

where $(1-\alpha_1)$ is called the differential capture fraction of the grid, owing to the fact that the grid may be at a positive potential (see Fig. 1). Also, a new fluctuation N' arises from the capture of electrons by the grid. The fluctuation N' is introduced above to include the effect of partition noise which will be considered later.

To find the total fluctuation in the plate current, it is necessary to know the value of I_p. The problem now is divided into two sections, one on the negative grid case and the other on the positive grid case.

(a) Negative Grid Case

In the case of a tube with negative grid, the value of α_1 above is equal to unity and $N'=0$. From Eq. (2) and Eq. (3) we obtain

$$\begin{aligned} I_p = a_{11}^{(2)}(V_p-V_g)+\{a_{12}^{(2)}a_{21}^{(1)}+a_{13}^{(2)}a_{31}^{(1)}\} \\ \times(V_g-V_a)+(v_f)_a\{a_{12}^{(2)}a_{23}^{(1)}+a_{13}^{(2)}a_{33}^{(1)}\}. \end{aligned} \tag{4}$$

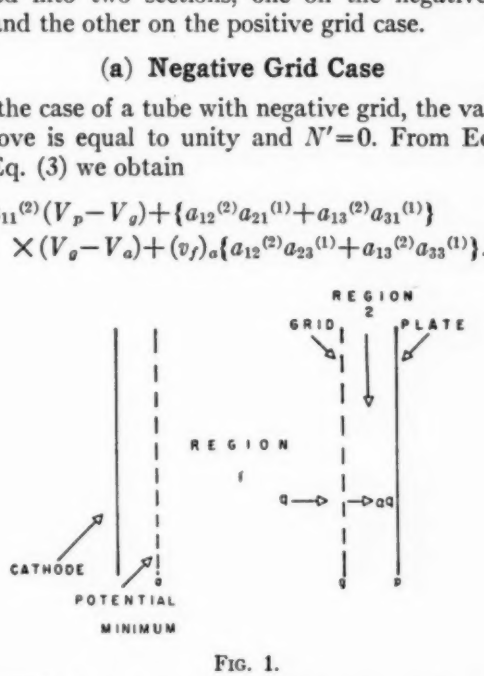


FIG. 1.

‖ For definition of the space-charge factor, see, for instance, reference 1.

¶ Superscripts of the coefficients $a_{11}, a_{12}\cdots a_{33}$ indicate the region concerned.

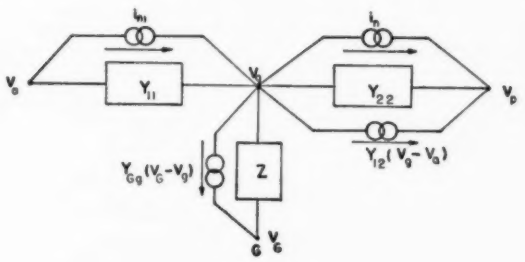


FIG. 2.

One can then draw the general equivalent circuit for a triode with a finite μ on a node basis as shown in Fig. 2, where V_G is the potential of the grid wires and Z is an impedance element connected between the actual grid wires and the grid plane in the electronic analysis. Z reduces to a pure capacitance when the grid is at a negative potential. The value of the capacitance agrees with that given by Maxwell's screening formula in the absence of electrons.[7] Other methods of calculating the equivalent grid potential have been used by North[12] and Zuhrt.[13] If the tube has an infinite μ (infinite amplification factor), i.e., if the regions (1) and (2) are completely shielded from each other, then the branch composed of Z and the current generator disappears from the above equivalent circuit.

In Fig. 2, Y_{22} is the self-admittance of region (2), Y_{12} is the mutual admittance between regions (1) and (2), and Y_{11} is the self-admittance of region (1). That is:

$$\begin{aligned} Y_{22} &= a_{11}^{(2)} \\ Y_{12} &= \{a_{12}^{(2)}a_{21}^{(1)}+a_{13}^{(2)}a_{31}^{(1)}\} \\ Y_{11} &= a_{11}^{(1)}. \end{aligned} \tag{5}$$

The corresponding value of noise current in the plate circuit owing to a fluctuation in velocity $(v_f)_a$ at the plane a is i_n given by

$$i_n=(v_f)_a\{a_{12}^{(2)}a_{23}^{(1)}+a_{13}^{(2)}a_{33}^{(1)}\}. \tag{6}$$

As is usually done, we assume that the region (2) is space-charge free, and that $\xi_2\to 0$. Hence,

$$i_n\to(v_f)_a(a_{12}^{(2)}a_{23}^{(1)})$$

and

$$a_{12}^{(2)}\cong 1\cdots \quad \text{for} \quad |\beta_2|\ll 1. \tag{7}$$

Where $\beta_2=j\theta_2=j\omega T_2$ for region (2), θ_2 is the transit angle of region (2), T_2 being the transit time of region (2). Substituting now the above result, we obtain for noise current in the plate circuit, $i_n\cong a_{23}^{(1)}(v_f)_a$, using the proper value of dc current in $a_{23}^{(1)}$ that is

$$\langle i_n^2\rangle_{Av}\cong|a_{23}^{(1)}|^2\langle(v_f)_a\rangle_{Av}. \tag{8}$$

One now uses the noise velocity of Rack[6] given by

$$\langle(v_f)^2_a\rangle_{Av}=(4\eta kT_c/eI_0)(1-\pi/4)\Delta f, \tag{9}$$

[12] D. O. North, Proc. Inst. Radio Engrs. **24**, 108 (1936).
[13] H. Zuhrt, Hochfreq. u. elek. **29**, 73 (1937).

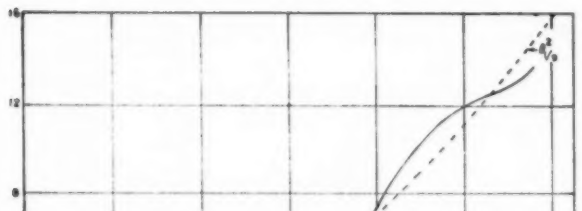


FIG. 3.

writing**

$$g_m = 1/r_p = (12\epsilon^2/I_0\eta T_1^4), \tag{10}$$

where $\eta = e/m$ for an electron, I_0 = cathode current, T_c = temp. of cathode, k = Boltzmann constant, Δf = band width of measuring apparatus, ϵ = dielectric constant of free space, and T_1 = transit time of region (1). We obtain,

$$\langle i_n^2\rangle_{Av} = S(\theta_1)[3(1-\pi/4)4kT_c g_m \Delta f]. \tag{11}$$

A plot of $S(\theta_1)$ occurring in Eq. (11) is made in Fig. 3 (reproduced from a private communication from J. R. Whinnery).

The following comments regarding the noise current generator in the input circuit are in order:

$i_{n1} = a_{13}^{(1)}(v_f)_a$, gives the noise current at the grid due to a velocity fluctuation at plane a.

$$\langle i_{n1}^2\rangle_{Av} = |a_{13}^{(1)}|^2\langle v_{fa}^2\rangle_{Av} = 3(1-\pi/4)4kT_c g_{10}\Delta f G(\theta_1) \tag{12}$$

where

$$G(\theta_1) = (4/\theta_1^3)[2+\theta_1^2-2(\cos\theta_1+\theta_1\sin\theta_1)] \tag{13}$$

and g_{10} = dc conductance of region (1). $G(\theta_1)$ gives the variation of noise current in the diode region with transit angle θ_1. In Eq. (12) one should use $\xi_1 = 1$ to correspond to the case of a diode or the cathode to grid region in a triode.

Evaluation of $\langle i_n^2\rangle_{Av}$ for Small Values of θ_1

$$a_{12}^{(2)} \cong 1 \cdots \text{ for } |\beta_2| \ll 1.$$

We have to evaluate $a_{23}^{(1)}$ for small values of θ_1

$$a_{23}^{(1)} \cong (2\epsilon/\eta T_1^2)[\beta_1 e^{-\beta_1} + 3] \tag{14}$$

neglecting terms involving powers ≥ 2 of θ_1. Then,

$$\langle i_n^2\rangle_{Av} \cong \langle (v_f)_a^2\rangle_{Av}(4\epsilon^2/\eta^2 T_1^4)(9+\theta_1^2). \tag{15}$$

Writing $\Gamma^2 = (3/eI_0)(2kT_c g_m(1-\pi/4))$ and using Eq. (9) and Eq. (10), Eq. (15) reduces to

$$\langle i_n^2\rangle_{Av} \cong 2eI_0\Delta f\Gamma^2(1+\theta_1^2/g) \tag{16}$$

for

$$\theta_1 \to 0, \quad \langle i_n^2\rangle_{Av} \cong \Gamma^2 2eI_0\Delta f, \tag{17}$$

that is, the limiting value of $\langle i_n^2\rangle_{Av}$ for small values of θ_1 is the same as in the case of a diode whose transit angle is negligible. This is predictable anyway since the effect of the grid-to-plate region was completely neglected.

Evaluation of $\langle i_n^2\rangle_{Av}$ for Large Values of θ_1

As in the previous case for $|\beta_2| \ll 1$, we obtain

$$i_n \cong a_{23}^{(1)}(v_f)_a$$

$$a_{23}^{(1)} \cong (2\epsilon/\eta T_1^2)\beta_1 e^{-\beta_1}\cdots \quad \text{for} \quad |\beta_1| \gg 1. \tag{18}$$

As before we write expressions for $\langle (v_f)_a^2\rangle_{Av}$, g_m, and Γ^2; after substitution and rearrangement of terms, we obtain

$$\langle i_n^2\rangle_{Av} \cong 2eI_0\Delta f\Gamma^2 F(\theta_1) \tag{19}$$

where $F(\theta_1) = \theta_1^2/9$. A plot of $F(\theta_1)$ of Eq. (19) is shown in Fig. 3 by a dotted line. Hence, the asymptotic behavior is that the noise power in the plate circuit of a triode varies as the square of the cathode-to-grid transit angle, provided the transit angle between the grid and the plate is small in comparison with the cathode-to-grid transit angle.

(b) Positive Grid Case: Partition Noise††

We use the same Llewellyn-Peterson equations for the equivalent diodes representing the cathode-to-grid and grid-to-plate regions in the following analysis.

$$I_p = a_{11}^{(2)}(V_p - V_g) + \{a_{12}^{(2)}\alpha_1 a_{21}^{(1)} + a_{13}^{(2)}a_{31}^{(1)}\} \times (V_g - V_a) + (v_f)_a\{a_{12}^{(2)}a_{23}^{(1)}\alpha_1 + a_{13}^{(2)}a_{33}^{(1)}\} + a_{12}^{(2)}N'. \tag{20}$$

As has been stated, the fluctuation N' is introduced in Eq. (20) to include the effect of partition noise, and we recall that the excess of noise in an electrode over the simple apportioning of cathode-current fluctuations is termed partition noise.

As before, concerning ourselves with the case $|\beta_2| \ll 1$ and $|\beta_1| \gg 1$, we obtain

$$i_n \cong \{\alpha_1 a_{23}^{(1)}(v_f)_a + N'\}$$

or

$$\langle i_n^2\rangle_{Av} \cong \alpha_1^2\Gamma^2 2eI_0\Delta f(\theta_1^2/9) + \langle (N')^2\rangle_{Av}, \tag{21}$$

where Γ^2 and other quantities have the same significance as in the negative grid case. N' is obtained from the expression‡‡

$$\langle (N')^2\rangle_{Av} = 2e\alpha_1(1-\alpha_1)I_D, \tag{22}$$

** Assuming initial average velocity is small at the potential minimum, the low-frequency plate impedance for a diode (in our case, the region between cathode and grid) can be written as shown.

†† The effect of secondary electrons is not considered in this paper. However, this is under investigation at present.

‡‡ This expression is obtained on the basis that the effect of secondary electrons can be neglected, and that the grid is so fine that the probability of an electron hitting it is the same, regardless of the position on the cathode from which it leaves.

where I_D is the dc current in the region preceding the grid. It can be seen above that as $\alpha_1 \to 1$, $N' \to 0$, meaning that under the above circumstances, the effect of positive grid is negligible.

III. NOISE CONVECTION CURRENT IN A DIODE AND NOISE IN ELECTRON STREAMS

One can see from Eq. (2) that the component of noise convection current at the exit plane of a diode region due to a fluctuation $(v_f)_a$ in the velocity at the entrance plane is given by

$$q_g = a_{23}^{(1)}(v_f)_a. \tag{23}$$

Hence, one can say that any investigation regarding the calculation of noise current in a triode under the conditions $\theta_1 \gg 1$, and $\theta_2 \ll 1$, and the experimental verification of the calculated result is one involving the noise convection current in a diode due to an initial velocity fluctuation at the entrance plane. Then, because of the above reason, we say that the present investigation is concerned with the measurement of noise in electron streams.

IV. EXPERIMENTAL TECHNIQUE AND RESULTS

For the present investigation the experimental tubes were designed with the following characteristics in mind:

(a) large transit angle of the cathode-to-grid region,
(b) small transit angle of the grid-to-plate region, and
(c) rectilinear flow of electrons (in the present case the beam is cylindrical). This is essential since it is desired to verify the results obtained from the one-dimensional theory of Llewellyn and Peterson.

Transit angle of any region can be varied at any frequency by varying the distance between the corresponding electrodes, other conditions remaining the same. It can also be seen that it is convenient to make measurements at a low enough frequency to reduce the problems of circuitry; but at the same time the frequency should not be too low since the tube becomes unusually large, and one would have to face problems regarding emission and maximum current that can be drawn to ensure space-charge limited operation in the cathode-to-grid region. An electrode configuration that was chosen to approach the desired characteristics is shown in Fig. 4. It may be mentioned here that even though in the experimental tubes of the present investigation an axial beam of finite lateral extent is produced by a Pierce[14] type electrode structure, such an approximation is valid considering the low-current densities at which the tube was operated to obtain large transit angles.

The transit angle is a measure of the importance of space-charge waves. For the type of model chosen, the space-charge noise and velocity waves are taken into account properly through the use of Llewellyn-Peterson equations.

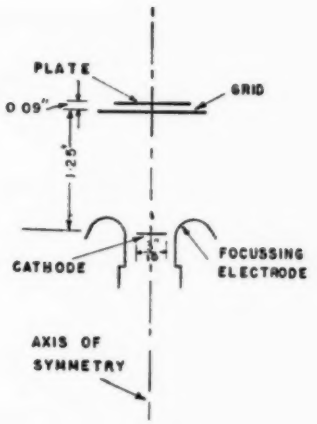


FIG. 4.

Also, considerable difficulty regarding the operation of initial experimental tubes has been experienced due to charging of the glass walls, which was circumvented later by using a better focusing electrode and a larger diameter glass envelope.

To obtain the variation of γ^2§§ with transit angle, one can make measurements either at a number of frequencies, or with various values of dc current flowing through the tube at a fixed frequency. In the present investigation the second alternative was chosen, except that measurements were made at two different frequencies, 34.6 mc/sec and 69.2 mc/sec. To measure γ^2 the noise generated in the experimental tube is compared with the noise from a saturated diode for which the transit time is negligible. It may be mentioned here that the external cathode-grid impedance is very small and consequently can be neglected, since the tube under consideration has a very low transadmittance which can be readily checked by calculation. One important feature about the measuring system used in the present investigation is that it has rather narrow band widths; this is very desirable since, if one obtains any information regarding γ^2, it will be at a particular value of transit angle, or more precisely at the average transit angle over a small range of frequencies.

It can be seen from Section II of this paper that Γ^2 is a function of the operating conditions. Hence, any computations made in the determination of $F(\theta_1)$ involves a fair amount of algebra, since the various values of T_1, g_m, and Γ^2 would have to be computed. The experimental curve for $F(\theta_1)$ *vs* θ_1 is drawn on a db scale in Fig. 5 for a negative grid case. We now recall that as $\theta_1 \to 0$, the triode under investigation would behave like a diode whose transit angle is small. The absolute value of noise at low frequency (θ_1 small) has been measured which serves as a check point. Since only a segment (not the whole) of the curve of $F(\theta_1)$ *vs* θ_1 can be obtained by measurements at either of the two fre-

[14] J. R. Pierce, J. Appl. Phys. **11**, 548 (1940).

§§ $\gamma^2 = \Gamma^2 F(\theta_1)$.

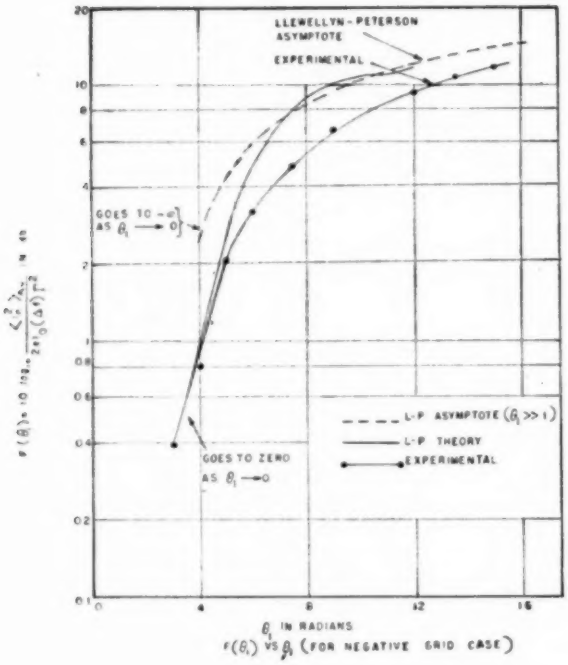


FIG. 5.

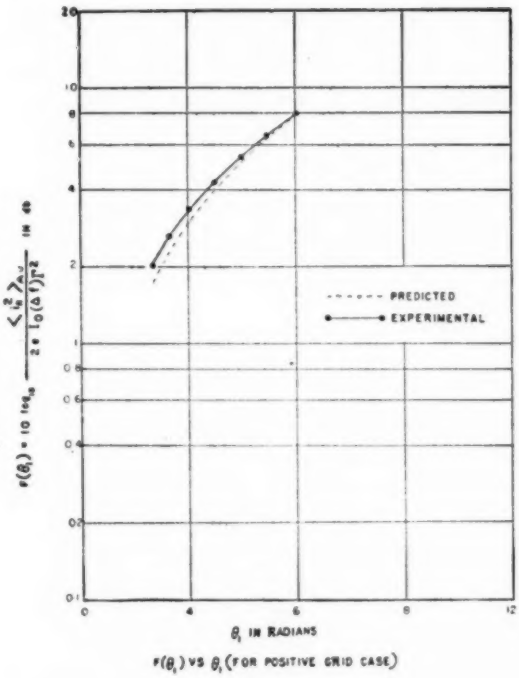


FIG. 6.

quencies, at least one check point is necessary to connect the two segments of the curve. In the present investigation, actually two such check points were obtained. It can be seen from Fig. 5 that the experimental curve approaches asymptotically a value which is about 2.0 db below the theoretical value predicted by the Rack approach.

Proceeding then to the positive grid||| case, one can write Eq. (21) in the following form:

$$\langle i_n^2\rangle_{Av}=2eI_0\Delta f\left[1-(1-\Gamma_1^2)\frac{I_0}{I_0+I_g}\right]=2eI_0\Delta f\Gamma^2F'(\theta_1),$$

where I_0+I_g is the cathode current, I_0 the plate current, I_g the grid current, and $\Gamma_1^2=\Gamma^2F(\theta_1)$. Hence, using the experimental values of $F(\theta_1)$ for the negative grid case one can obtain the predicted values of $F'(\theta_1)$ for the positive grid case. Figure 6 presents both the experimental and predicted curves.

Investigations have been made regarding the convergence and divergence of the beam by changing the potential of the focusing electrode. As expected, it was found that the noise in a converging beam is more, and in a diverging beam less, than in the case of a beam with rectilinear flow. Also, preliminary calculations made using the space-charge wave theory indicate that the finite diameter of the beam explains a reduction of noise of about 1 db at large transit angles ($\sim$12 radians), although similar reduction at lower transit angles is not explained. This is probably owing to the lack of validity of the assumptions made in space-charge wave theory under large current conditions.

V. CONCLUSION

It can be seen from the above that the general agreement between the theoretical predictions and experimental results is good. Hence, this proves that a single-valued velocity theory is powerful enough to predict many of the characteristics of a vacuum tube, under small signal conditions.

There is reason to believe that the main weakness of the theory is its assumption of single-valued velocity for the electrons. Hence, much work has to be done in this direction before the problem is completely understood. The modification of the Rack approach suggested by Robinson[15] does not seem to be applicable in the present investigation. It appears, however, that the finite diameter of the beam does play an important role in the foregoing study, and is being investigated at present in greater detail.

VI. ACKNOWLEDGMENTS

The writer is greatly indebted to Professor L. J. Black and Professor J. R. Whinnery for supervision and encouragement during the course of this work, and he wishes to thank Mr. F. D. Clapp for his assistance in the design of the measuring system, Mr. G. A. Becker for building the experimental tubes, and the Electronics Research Laboratory of the University of California, Berkeley, for extending its many facilities to him.

||| The value of θ_1 in the case when the grid is at a positive potential is computed by treating the cathode-to-grid region as a diode whose dc current is (I_0+I_g).

[15] F. N. H. Robinson, Phil. Mag. 43, 863 (1950).

Some Phenomena Associated with Supersonic Liquid Jets*

BRIAN DUNNE AND BENEDICT CASSEN
Atomic Energy Project, School of Medicine, University of California, Los Angeles, California
(Received August 3, 1953)

Supersonic liquid jets were produced in air by means of a spring-loaded injector. At high jet velocities another type of breakup seems to occur besides the classical Rayleigh surface tension breakup, and the sinuous aerodynamic breakup. Rotationally symmetric waves are formed, and appear to "break" analogous to wind-produced waves on a body of water. The rate of cavitation of jets in water was studied for several jet velocities above the speed of sound in air.

INTRODUCTION

As early as 1878, Rayleigh[1] discussed the instability of liquid jets. In his treatment, the jet was assumed to be a cylinder of nonviscous fluid along which rotationally symmetric waves could be propagated by surface tension restoring forces. Since the surface area of the jet is not minimal for the cylindrical form, but for droplets of a certain radius, the cylindrical form is unstable. Rayleigh found that if the wavelength of a disturbance was greater than the circumference of the jet, the amplitude of the disturbance increased exponentially with time, and that the wavelength of maximum instability was 1.4 times the circumference.

Weber[2] and Haenlein[3] studied another type of breakup in jets which occurs at velocities on the order of 80 m/sec, where air friction forces become more pronounced. The effect of air friction was to modify the Rayleigh breakup, and also give rise to unstable sinuous vibrations of the liquid column. Since parts of the jet which are at greater radial distances from the original column axis have aerodynamic forces which tend to increase this distance, the oscillation amplitude increases until the jet is sheared.

With velocities close to and exceeding the speed of sound in air, it appears that another type of breakup may occur. This phenomenon is the subject of the first part of this report. During the recent war, considerable interest was aroused in the theory of liquid jets by the development of metal-shaped charge jets produced by conical liners in high explosive cavities.[4] Because of the high velocities obtained (about 10^4 m/sec) the interaction of a jet with target material was assumed, in the simple theory, to be comparable to a liquid jet striking a liquid target. In this sense, the second part of this article can be considered somewhat analogous to the metal jet penetration of metal targets.

EXPERIMENTAL PROCEDURE

The high-speed jets were produced by an injector used in biological applications which is manufactured by the R. P. Scherer Corporation, of Detroit, Michigan. The mechanism is essentially a spring-loaded plunger which forces a rubber stopper into one end of a small, drawn, stainless-steel capsule, the other end having a small hole, through which the jet emerges. The cross sections of the two nozzle shapes used in these experiments are shown in Fig. 1(a); the two hole diameters used were 0.13 mm and 0.20 mm.

The photographic equipment consisted of two 0.25 microsecond-duration Libessart sparks with a variable electronic delay between them consisting of a multivibrator circuit. The experimental arrangement is shown in Fig. 1(b). The jet was shot vertically downward into a small box which contained a photographic plate and a variable width slit facing the spark sources. This arrangement allowed two successive frames to be taken on the same photographic plate, from 10 to 120 microseconds apart.

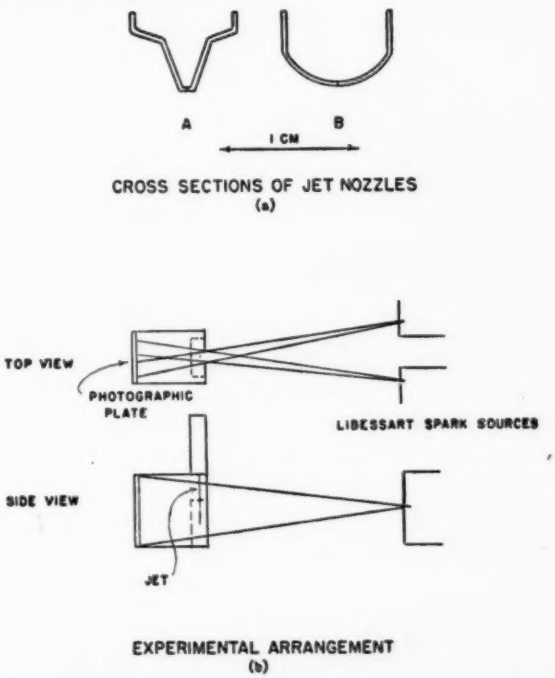


FIG. 1.

* This paper is based on work performed under Contract No. AT-04-1-GEN-12 between the U. S. Atomic Energy Commission and the University of California at Los Angeles.

[1] Lord Rayleigh, Proc. Lond. Math. Soc. **10**, 7 (1878).
[2] C. Weber, Z. angew. Math. u. Mech. **2**, No. 3 (1931).
[3] A. Haenlein, *Forshung Gebiete Ingenierw.* **2**, No. 4 (1931). Natl. Advisory Comm. Aeronaut. Tech. Mem. 659.
[4] Birkhoff, MacDougall, Pugh, and Taylor, J. Appl. Phys. **19**, 563 (1948).

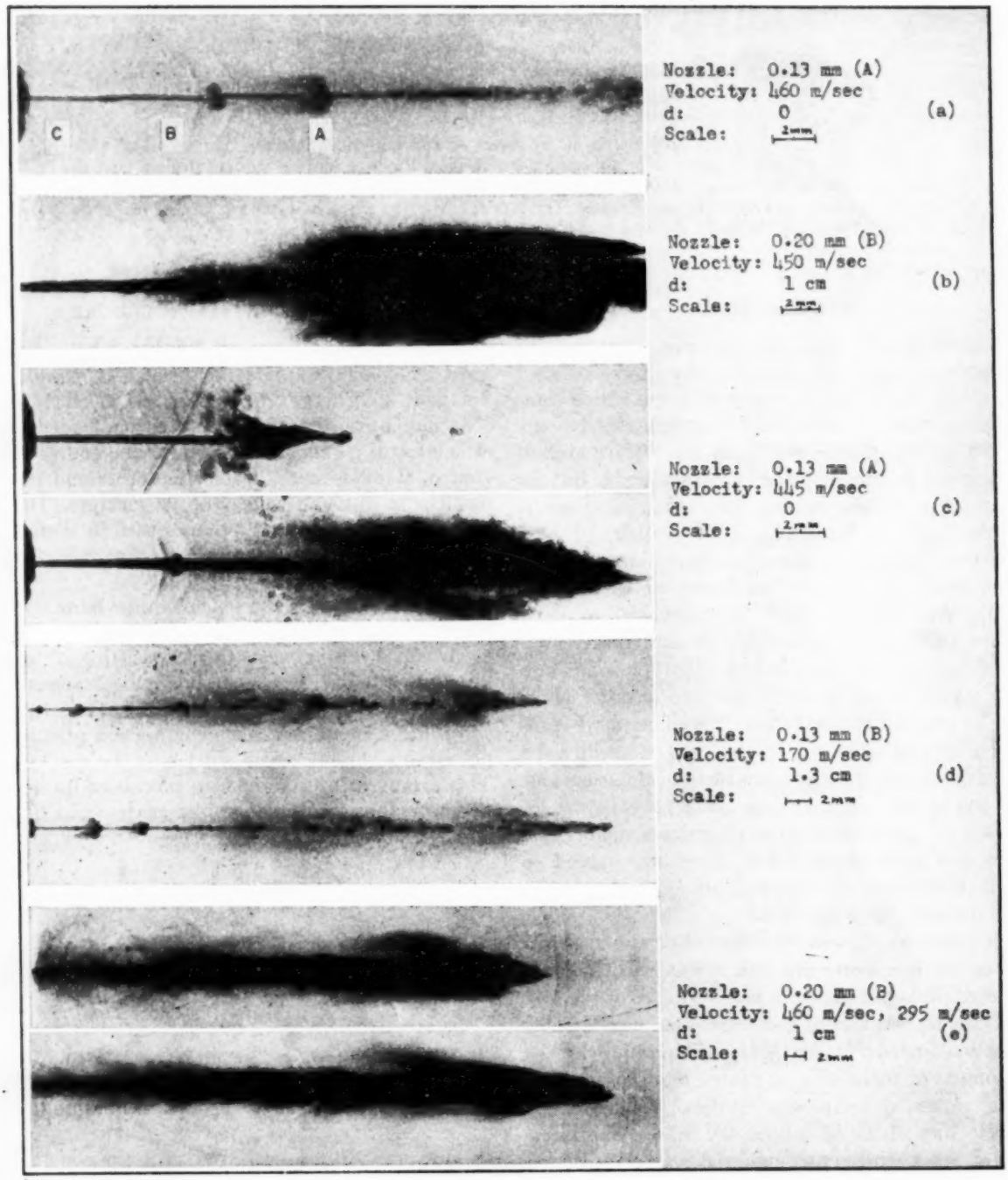


FIG. 2. Supersonic and subsonic jets. "*d*" is the distance from the left-hand side of the photograph to the nozzle.

In order to increase the opacity of the jet for the shadowgraph technique, the fluid used was an aqueous solution of methylene blue (2.4 grams per liter) with a surface tension approximately ten percent higher than that of distilled water. There was no apparent difference in behavior between jets of distilled water and jets of the methylene blue solution, however. The jets were injected into air at atmospheric pressure.

SUPERSONIC JETS IN AIR

Figure 2(a) is a shadowgraph of a jet from a 0.13-mm Type A nozzle illustrating a type of breakup which sometimes occurs at high jet speeds. "*A*" is the leading wave accompanied by a detached head and a tail shock wave. "*B*" is another wave which appears to be partially shielded aerodynamically by "*A*." At "*C*," a third wave is just beginning to form. The wave at "*A*" seems

to be "breaking" analogously to waves on an open body of water under high wind conditions. The detached head shock waves of waves "*A*" and "*B*" appear to be weak in the vicinity of the jet axis, indicating the possibility of a boundary layer near the jet, which could be caused by the initial part of the jet accelerating the air in this region. The speed of this jet as measured by the shock wave angles is about 460 m/sec. The spray trail from wave "*A*" was observed on the negative to extend back as far as wave "*B*." The breaking waves are apparently not caused by the initial pressure transients alone, since they have been observed in later parts of jets in several photographs, of which Fig. 2(b) is an example. This jet is from a 0.20-mm Type B nozzle, and has a velocity of approximately 450 m/sec. The Reynolds number at the jet orifice is about 9×10^4; on this basis the fluid might be assumed to be highly unstable to turbulent flow.

The second frame of Fig. 2(c) shows another wave accompanied by head and tail shock waves. This jet, from a 0.13-mm Type A nozzle, is accelerating, since there are violent collisions evidenced in the second frame, and there are no shock waves visible in the first frame; the diagonal line on the first frame of Fig. 2(c) is a scratch on the negative. The velocity of the wave is approximately 445 m/sec.

The breaking waves can be seen on subsonic jets, but the form of the wave is not so definite as in the supersonic case. Figure 2(d) shows two successive frames of a subsonic jet from a 0.13-mm Type B nozzle with a velocity of approximately 170 m/sec. From this photograph, it appears that at this velocity, the wavelength of maximum instability is about five times the jet circumference.

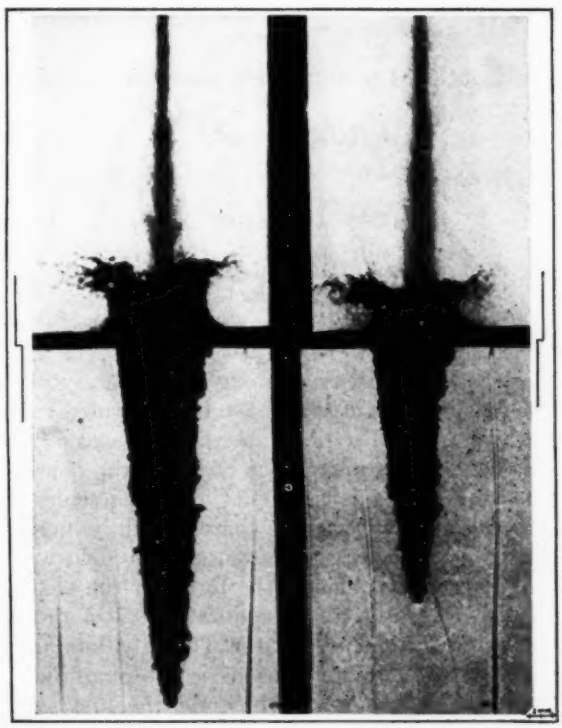


FIG. 3. Supersonic jet cavitation in water.

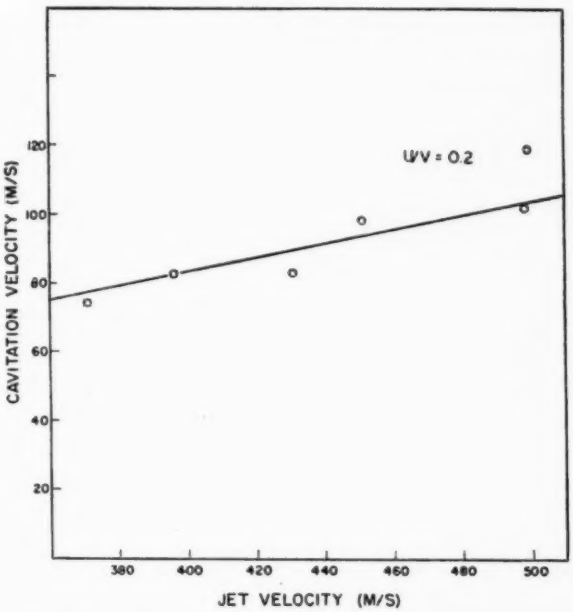


FIG. 4. Jet cavitation velocity (U) *versus* jet velocity (V).

Figure 2(e) shows a supersonic jet from a Type B 0.20 mm nozzle. The jet tip has two head shock waves, a phenomenon which has been observed in several photographs. The appearance of the two non-stationary shock waves may be due to abrupt changes in frontal area as the jet sheds. The velocity of the jet near the orifice (as measured by free-stream shock waves) is about 460 m/sec, while the tip velocity is 295 m/sec, the ratio of tip velocity to jet velocity being 0.64.

CAVITATION PRODUCED BY SUPERSONIC JETS

The jets were studied with regard to their interaction with a surface of a liquid to determine their velocity of cavitation. In this study, the jets were shot vertically downward into a glass-walled container 8 cm by 5 cm, and 10 cm high, filled with water. The position of this container is shown by the dotted lines in Fig. 1(b). The distance from the jet nozzle to the water surface was 8 cm. Two successive shadowgraphs with an interval of 78 microseconds between them are shown in Fig. 3. The jet velocity at a distance of approximately 5 cm from the water surface can be determined from the free-stream shock wave angle; and the speed of cavitation by measuring the cavity-tip displacement, since the time interval between the shadowgraphs is known.

The vertical lines in Fig. 3 are produced by sugar crystals dropped into the water just before the shadowgraph was taken, on a line parallel to the photographic plate, and passing through the jet, in order to observe

the displacement of the water due to the cavitation. A small shock wave in the water is visible in the right-hand photograph of Fig. 3. The lines drawn on each side of Fig. 3 indicate how the cavity is displaced from its true position by refraction. The scale shown at the bottom of Fig. 3 is nearly the same for both the top and bottom parts of the photograph, however.

The velocity of cavitation U is plotted against the jet velocity V, for several supersonic jet velocities in Fig. 4. These values were determined from cavity depths varying between 3 and 5 cm, where a steady state cavitation rate might be considered to prevail.

ACKNOWLEDGMENT

The authors wish to express their appreciation to Mr. Herbert Gass, who constructed most of the experimental equipment, and who assisted materially with the experiments.

JOURNAL OF APPLIED PHYSICS VOLUME 25, NUMBER 5 MAY, 1954

Transverse Motion of an Electron in a Constant Wave Speed Section of a Linear Accelerator*

IRA JACOBS AND E. S. AKELEY
Purdue University, West Lafayette, Indiana
(Received August 10, 1953)

The relativistic differential equations of motion of an electron in the field of a linear accelerator are considered. Treating only electrons near the axis, the field components are expanded in powers of distance from the axis r, and terms in r and $\dot{r}$ of order two higher than the lowest order terms are neglected. The first integral of two of the equations of motion are then obtained, and together with the equation of motion for the r coordinate, give a second-order nonlinear differential equation for r as a function of the phase of the electron with respect to the peak of the accelerating electric field. Utilizing earlier results giving the dependence of phase on distance along the axis of the accelerator, a choice of a particular dc magnetic focusing field allows this differential equation to be integrated twice, from which r as a function of distance along the accelerator is plotted.

The need for a focusing field is indicated, and it is shown that the particular field chosen has the advantage that with it, the electron will oscillate between known fixed limits of r. The necessity of shielding the electron emitter from the focusing field is shown, and an indication of the magnitude of the field and length of the accelerator for which it should be used, is given for the Purdue accelerator.

INTRODUCTION

COMPLETE electron orbits for the first two sections of the Purdue linear electron accelerator have been worked out by Caplan and Akeley[1] for electrons injected along the axis of the accelerator. For the axial electron, there is only a longitudinal force, and the problem is one-dimensional. In practice, however, all of the electrons will not be injected exactly along the axis, and the electron will experience transverse defocusing forces. To counteract these defocusing forces, a dc focusing magnetic field is used over part of the length of the accelerator.

It is the purpose of this work to investigate the transverse motion of the electron under the action of the combined electromagnetic and focusing fields.[2]

INTEGRATION OF THE EQUATIONS OF MOTION

We will consider an axially symmetric TM wave upon which is superimposed an axially symmetric dc magnetic field. The total field components are then given by

$$E_z = E_0 J_0(k_r r)\cos\psi,^{3}$$
$$E_\theta = 0,$$
$$E_r = (k_r v)^{-1} E_0 J_1(k_r r)\sin\psi,$$
$$H_z = H_z(r,z),$$
$$H_\theta = k_r^{-1} E_0 J_1(k_r r)\sin\psi,$$
$$H_r = H_r(r,z).$$

In these equations, distance along the axis and distance off the axis in units of $\lambda/2\pi$ have been denoted respectively by z and r. The phase is expressed as $\psi = z/v - t$, where the time t is in units of $1/\omega$, and the phase velocity of the wave v, is in units of c. For convenience in the discussion, the electronic charge e will be taken as positive, and the direction of the electric field will be defined as the direction of the force on the electron. $k_r = (1-v^{-2})^{\frac{1}{2}}$, and is imaginary for $v = 0.7$ in the first section of the Purdue accelerator. The only dimensional quantity in the above equations is the amplitude E_0 of the axial electric field, which for the first section

* Published as Linear Electron Accelerator Technical Report No. 1, Department of Physics, Purdue University, June, 1952.

[1] D. Caplan and E. Akeley, J. Appl. Phys. **24**, 774 (1952).

[2] E. L. Chu, The Theory of Linear Electron Accelerators, M. I. Report No. 140 (Microwave Laboratory, Stanford University, May, 1951). Chapter six deals with the subject of this paper in a different manner.

[3] J. C. Slater, *Microwave Electronics* (D. Van Nostrand Company, Inc., New York, 1950), p. 276.

of the Purdue accelerator is 110 statvolts/cm. The dc focusing field will as yet be kept general.

The relativistic differential equations of motion in cylindrical coordinates are then found to be

$$\frac{d}{dt}(m\dot{r})-mr\dot{\theta}^2-\frac{e}{c\omega}\frac{E_0}{k_r}J_1(k_r r)\left(\frac{1}{v}-\dot{z}\right)\sin\psi - \frac{e}{c\omega}r\dot{\theta}H_z=0, \quad (1a)$$

$$\frac{d}{dt}(mr^2\dot{\theta})-\frac{e}{c\omega}r(\dot{z}H_r-\dot{r}H_z)=0 \quad (1b)$$

$$\frac{d}{dt}(m\dot{z})-\frac{e}{c\omega}\left\{\frac{\dot{r}}{k_r}E_0J_1(k_r r)\sin\psi-r\dot{\theta}H_r\right\} - \frac{eE_0}{c\omega}J_0(k_r r)\cos\psi=0. \quad (1c)$$

If H_z is constant, then $H_r=0$, and the integration of (1b) then yields

$$mr^2\dot{\theta}=-er^2H_z/2\omega c+C. \quad (2)$$

If we consider a more general H_r and H_z, and expand these fields in powers of r, Maxwell's equations, together with the neglect of second and higher order terms, leads directly to the same result (2). Thus, if the electron is sufficiently close to the axis, the result (2) can be used even if H_z is not constant, and $\dot{\theta}$ may be eliminated from Eqs. (1a) and (1c), giving

$$\frac{d}{dt}(m\dot{r})-\frac{C^2}{mr^3}+\frac{r}{4m}\left(\frac{eH_z}{\omega c}\right)^2 - \frac{e}{\omega c}\frac{E_0}{k_r}J_1(k_r r)\left(\frac{1}{v}-\dot{z}\right)\sin\psi=0, \quad (3a)$$

$$\frac{d}{dt}(m\dot{z})=\frac{e}{\omega c}\frac{E_0}{k_r}\dot{r}J_1(k_r r)\sin\psi-\frac{er}{mc\omega}H_r\left(\frac{C}{r^2}-\frac{eH_z}{2\omega c}\right) - \frac{eE_0}{c\omega}J_0(k_r r)\cos\psi=0. \quad (3b)$$

It is to be noted in Eq. (3a) that the C^2/mr^3 term corresponds to a defocusing force which increases as the particle approaches the axis. It would then be desirable to make $C=0$. If the subscript c refers to conditions at the cathode of the electron gun, from (2)

$$C=(m_c\dot{\theta}_c+eH_{z,c}/2\omega c)r_c^2.$$

Since $\dot{\theta}_c$ is a statistical quantity, it cannot be hoped to have $C=0$ for all particles. The best that can be done is to make $H_{z,c}=0$, i.e., to shield the cathode from the magnetic field. The effect of not shielding the cathode will be shown more clearly by the graphs of the motion.

Neglecting second and higher order terms in r and $\dot{r}$, Eq. (3b) reduces to the equation of motion of an axial particle, the first integral of which is readily obtained as

$$(1+p_z^2)^{\frac{1}{2}}/v-p_z=A(\sin\psi+k), \quad (4)$$

where $p_z=\dot{z}/(1-\dot{z}^2)^{\frac{1}{2}}$, $A=eE_0/m_0c\omega$, and k is an integration constant to be determined by the initial conditions. Thus the axial longitudinal solution may be assumed to be unchanged for particles with r and $\dot{r}$ sufficiently small so that second-order terms may be neglected in comparison to the other terms in Eq. (3b).

Using Eq. (4) and neglecting third-order terms, Eq. (3a) may then be written in the form

$$\frac{d}{dt}(m\dot{r})=C^2/mr^3-(m_0A^2/4)r(1-\dot{z}^2)^{\frac{1}{2}} \times\{(H_z/E_0)^2-2\sin\psi(\sin\psi+k)\}. \quad (5)$$

[Since $\dot{z}$ is implicitly given as a function of ψ by Eq. (4), the time derivatives may be changed to derivatives with respect to ψ, and the variables $\dot{z}$ and t may be eliminated from Eq. (5). This results in a nonlinear second-order differential equation. If $C=0$, the equation is linear and may be put in the form of the general Lamé equation, and a series solution may be obtained. We, however, wish to consider the effect of the nonlinear term, and we also desire solutions which can be used for computation other than in the immediate vicinity of the point about which the expansion is made.]

As was mentioned before, the term C^2/mr^3 which is due to the initial angular momentum of field plus particle, corresponds to a defocusing force. This force becomes very large as the particle approaches the axis, so that if $C=0$, the electron will never return to the axis. This defocusing force becomes less appreciable, however, as the energy of the particle increases, because of the inverse dependence on m. Equation (5) also shows that the sense of the H_z field is immaterial.

In making rough calculations for the magnitude of the focusing field, Chu[2] assumes C to be zero, and takes as the focusing condition that $d/dt(m\dot{r})\leqslant 0$. (This is a stronger condition than just taking $\dot{r}$ to be decreasing, since m itself increases in the motion.) This calculation may be easily made from Eq. (5), if again C is taken to be zero. Then for $d/dt(m\dot{r})\leqslant 0$, $(H_z/E_0)^2\geqslant 2\sin\psi(\sin\psi+k)$. From Caplan's and Akeley's[1] results, $2(\sin\psi)(\sin\psi+k)$ is plotted against z (Fig. 1) for particles having initial phase in the first quadrant. It is apparent from Eq. (5) that a positive value of $2\sin\psi(\sin\psi+k)$ corresponds to a defocusing force. From Fig. 1, it is seen that for particles having an initial phase in the range between 60° and 90°, $2\sin\psi(\sin\psi+k)$ is always positive; and for particles with initial phase between 0° and 60°, $2\sin\psi\times(\sin\psi+k)$ is positive at the beginning of the accelerator. Therefore, at the beginning of the accelerator where the focusing problem is most acute (because of a relatively small $\dot{z}$); for the initial phases that give the

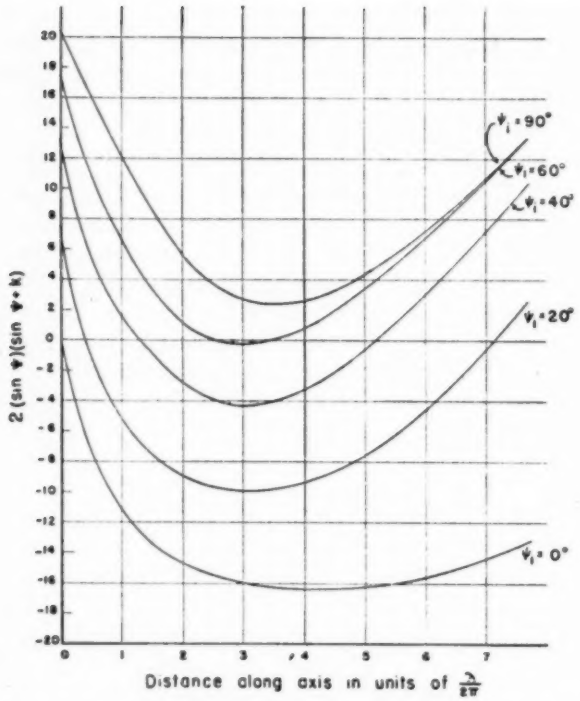


FIG. 1. $2\sin\psi(\sin\psi+k)$ plotted as a function of z for initial phases in the first quadrant, as calculated from the results of Caplan and Akeley (reference 1).

largest energies, the radial force is defocusing. This is the reason for the use of a focusing magnetic field.

From Fig. 1, it is seen that the largest value of $2\sin\psi(\sin\psi+k)$ occurs at the injection of the particle with initial phase 90° and is equal to 20.47. In the first section of the Purdue accelerator, $E_0=110$ statvolts/cm. Then for $H_z \geqslant 498$ gauss, focusing action will occur throughout the entire section.

While a general solution of Eq. (5) will not be attempted, a solution valid throughout the entire first section, and showing the effect of an initial angular momentum ($C\neq 0$), will be obtained under a somewhat special assumption.

Let $\epsilon=(H_z/E_0)^2-2\sin\psi(\sin\psi+k)$, and then Eq. (5) may be rewritten as

$$\frac{d}{dt}(m\dot{r})=C^2/mr^3-(m_0A^2/4)\epsilon r(1-\dot{z}^2)^{\frac{1}{2}}. \qquad (6)$$

A general solution of Eq. (6) can and will be obtained under the special assumption that ϵ is constant. This will be approximately true if H_z is constant and sufficiently large so as to make the variation in $2\sin\psi \times(\sin\psi+k)$ negligible in comparison to $(H_z/E_0)^2$. However, that would necessitate a constant field of several thousand gauss which is unnecessary. Referring again to Fig. 1, it is to be noted that the curves for initial phase between 20° and 90° are similar in appearance. Thus if an H_z field were set up so as to make ϵ constant for one particle, then it would be essentially constant for all particles with initial phase between 20° and 90°.

The desirability of having ϵ constant is not only for mathematical convenience, but as the solutions will show, if ϵ is constant, then r will oscillate between fixed limits which can be calculated from the first integral. The field considered is shown in Fig. 2.

Taking ϵ constant, the integration of Eq. (6) gives

$$\dot{r}=\pm\{(1-\dot{z}^2)(\kappa-\epsilon A^2r^2/4-C^2/m_0^2r^2)\}^{\frac{1}{2}}, \qquad (7)$$

where κ is a constant of integration. While a general solution will be obtained, a knowledge of the motion may be obtained directly from the first integral (7).

Consider first $C=0$. Since ϵ is a positive constant, it follows from Eq. (6) that $\dot{r}$ is everywhere decreasing (except when the particle crosses the axis, in which case $\dot{r}$ changes sign.) Then, if at injection $\dot{r}_i>0$, r will increase until $\dot{r}$ becomes zero when $r=(2/A)(\kappa/\epsilon)^{\frac{1}{2}}$. Then r will begin to decrease, and will reach the axis with $\dot{r}=-\{\kappa(1-\dot{z}^2)\}^{\frac{1}{2}}$. As the particle crosses the axis, $\dot{r}$ changes to $+\{\kappa(1-\dot{z}^2)\}^{\frac{1}{2}}$ and again decreases to zero, so that the particle will oscillate about the axis with constant amplitude equal to $(2/A)(\kappa/\epsilon)^{\frac{1}{2}}$.

If $C=0$, there will be two different values of r at which $\dot{r}=0$, given by

$$r^2=2\kappa/\epsilon A^2\pm(2/\epsilon A^2)(\kappa^2-\epsilon A^2C^2/m_0^2)^{\frac{1}{2}}. \qquad (8)$$

From Eq. (6), the radial force acting on the electron is

$$(1-\dot{z}^2)^{\frac{1}{2}}(C^2/m_0r^3-\epsilon rm_0A^2/4), \qquad (9)$$

and it is to be remembered that a positive force is a defocusing force. Neglecting $\dot{\theta}_c$, $C=er_c^2H_{z,c}/2\omega c$ where

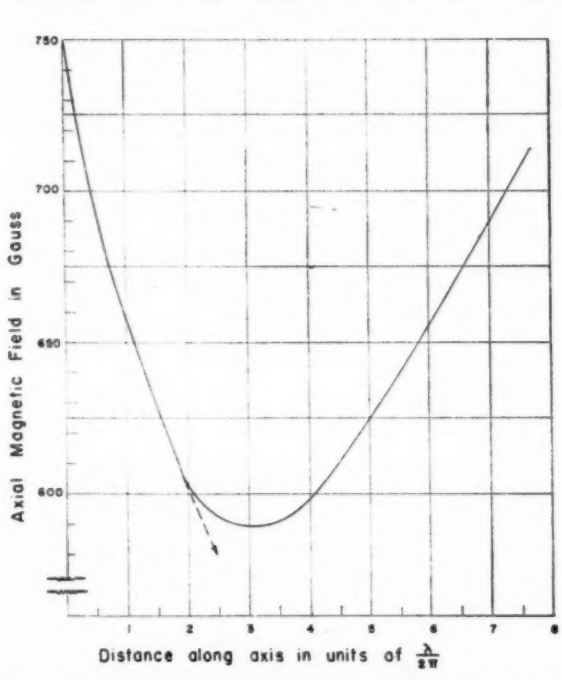


FIG. 2. Axial magnetic field as a function of distance along axis. The field is taken to be 750 gauss at the beginning of the accelerator, and the variation is chosen to make ϵ constant for the particle with initial phase 60°. (The variation in the values of ϵ for the particles with initial phase 20° and 90° are then less than 10 percent.)

again the subscript c refers to values at the cathode of the electron gun. Then at the time of injection

$$\left(\frac{C^2}{m_0r^3}-\frac{\epsilon r m_0 A^2}{4}\right)_i=\frac{e^2H_{z,c}{}^2r_i}{4m_0\omega^2c^2}\left\{\left(\frac{r_c}{r_i}\right)^4-\left(\frac{E_0}{H_{z,c}}\right)^2\left[\left(\frac{H_{z,i}}{E_0}\right)^2-2\sin\psi_i(\sin\psi_i+k)\right]\right\}. \quad (10)$$

Consider now the case where the cathode is not only unshielded, but in addition the field at the cathode is equal to the field at the beginning of the accelerator. Assume also that $r_c=r_i$, which will be true for particles travelling parallel to the axis of the electron gun. Then

$$\left(\frac{C^2}{m_0r^3}-\frac{\epsilon r m_0 A^2}{4}\right)_i=\frac{E_0{}^2r_ie^2\sin\psi_i(\sin\psi_i+k)}{2m_0\omega^2c^2}. \quad (11)$$

It is noted that under these conditions, the radial force on the electron at the beginning of the accelerator is a defocusing force. What is even more important, is that this radial force is independent of H_z, which most certainly defeats the purpose of the magnetic field. From Eq. (10) it is clear that for a given $H_{z,i}$, if $H_{z,c}=0$ (i.e. if the cathode is shielded), then the focusing force at the beginning of the accelerator will be a maximum.

While further deductions may be made from Eq. (6) and its first integral (7), the general solution will show much more clearly these deductions. From Eqs. (7) and (4), it follows that

$$\pm(\kappa-\epsilon A^2r^2/4-C^2/m_0{}^2r^2)^{-\frac{1}{2}}dr=\pm A^{-1}\{(\sin\psi+k)^2-g^2\}^{-\frac{1}{2}}d\psi, \quad (12)$$

where $g^2=A^{-2}(v^{-2}-1)$. In the expression on the right, the minus sign is used for $\dot{z}<v$ (before the phase minimum), and the positive sign for $\dot{z}>v$. The integral of this term is an elliptic integral of the first kind as may be seen by making the substitution $y=\sin\psi$.

Integration of Eq. (12) then gives

$$r^2=-\frac{2\kappa'}{\epsilon A^2}\sin\left[\left(\frac{\epsilon}{g}\right)^{\frac{1}{2}}\{sn^{-1}(S_{1i},S_2)-sn^{-1}(S_1,S_2)\}+\sin^{-1}\left(\frac{\kappa-\epsilon A^2r_i^2/2}{\kappa'}\right)\right]+\frac{2\kappa}{\epsilon A^2}, \quad (13)$$

and in the special case when $C=0$,

$$r=\frac{2}{A}\left(\frac{\kappa}{\epsilon}\right)^{\frac{1}{2}}\sin\left[-\frac{1}{2}\left(\frac{\epsilon}{g}\right)^{\frac{1}{2}}\{sn^{-1}(S_{1i},S_2)-sn^{-1}(S_1,S_2)\}+\sin^{-1}\left\{\frac{r_iA}{2}\left(\frac{\epsilon}{\kappa}\right)^{\frac{1}{2}}\right\}\right], \quad (14)$$

where

$$S_1=\left[\left(\frac{2g}{1+k-g}\right)\left(\frac{1-y}{y+k+g}\right)\right]^{\frac{1}{2}},$$

$$S_2=\frac{1}{2}\left[\frac{k^2-(g-1)^2}{g}\right]^{\frac{1}{2}}$$

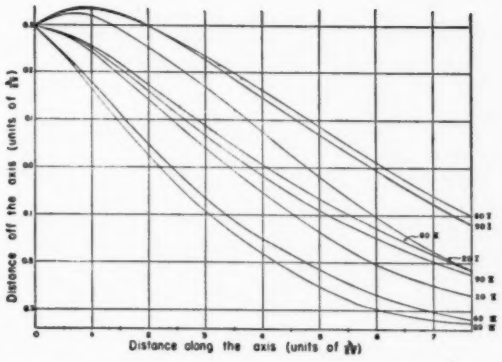


FIG. 3. r,z diagram for shielded cathode. The initial phase is labelled on each orbit. $r_i=0.296$ (=0.5 cm) for each orbit. Orbits labeled I, II, and III have $\dot{r}_i=0.1\dot{z}_i$, 0, $-0.1\dot{z}_i$, respectively.

and

$$\kappa'=(\kappa^2-\epsilon A^2C^2/m_0{}^2)^{\frac{1}{2}}.$$

Both of these solutions apply only before the phase singularity and before the first maximum in r. The modifications of the solution beyond these points are obtained by integrating in steps, taking into account sign changes in the functions.

CONCLUSIONS

Solutions (13) and (14) give r as a function of ψ. The dependence on z is then obtained from Caplan and Akeley's[1] results of phase as a function of distance. r is then plotted against z for the shielded cathode in Fig. 3, and for the unshielded cathode in Fig. 4. (For the unshielded cathode, C was evaluated assuming that the focusing magnetic field is constant in the electron gun, and that the electrons travel parallel to the axis of the gun.) For the Purdue accelerator, the electrons are injected with an initial energy of 50 kev, and it is assumed that no electron enters the accelerator more than 0.5 cm off the axis.

The solutions for the unshielded cathode (Fig. 4) indicate that at injection r is a minimum and defocusing will ensue. However, the actual solutions represented in Fig. 4 cannot be considered valid for the large values of r where the approximations made would not be

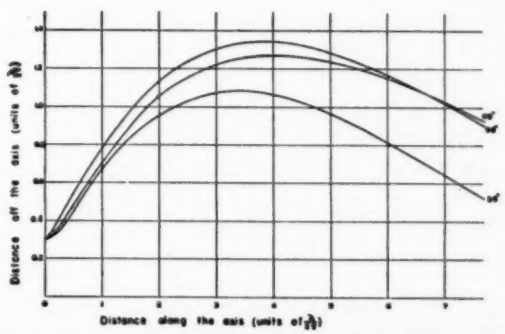


FIG. 4. r,z diagram for unshielded cathode. The magnetic field at the cathode is 750 gauss. $r_i=0.296$, $\dot{r}_i=0$. The initial phase is labeled on each orbit.

valid. However, the fact that r does become large is enough to make the situation undesirable and to be avoided. If the magnetic field at the cathode were less than that at the beginning of the accelerator, then the defocusing would not be as pronounced. However, as was indicated previously, for maximum focusing, the magnetic field at the cathode should be zero.

As the energy of the particle increases, the radial forces become small. For this reason it is unnecessary to have the focusing field over the entire first section. In Fig. 2, if the focusing field, instead of having a minimum at $z=3$, were to continue dropping from $z=2$ ($=3.37$ cm) as indicated by the broken line, then the focusing force for $z=2$ would be less than that giving the orbits of Fig. 3. However, at $z=2$ all the electrons with $r_i=0.296$ ($=0.5$ cm) will be approaching the axis, so that a decrease in the focusing force will mean a decrease in the slope of the curves, and consequently a larger value of z at which the electron will cross the axis. Preliminary calculations indicate that the values of r at the end of the first section of the accelerator would then be less than those indicated by Fig. 3. Thus it would appear that cutting off the field not only gives a saving in power, but will also improve the focusing.

If the maximum value of the field employed is less than 750 gauss (but still greater than 498 gauss), then the field should be used over a greater length of the section. However, since it was necessary to fit the magnetic field to make $\epsilon=(H_z/E_0)^2-2\sin\psi(\sin\psi+k)$ constant, smaller fields would introduce larger errors and make the results of this paper less applicable.

JOURNAL OF APPLIED PHYSICS VOLUME 25, NUMBER 5 MAY, 1954

Friction of Clean Metals and Oxides with Special Reference to Titanium

E. S. MACHLIN AND W. R. YANKEE
School of Mines, Columbia University, New York, New York
(Received September 17, 1953)

The friction coefficients of freshly cut surfaces in inert atmosphere are compared with values on the same surfaces later exposed to air. The kind of metal or oxide comprising the friction couple was also varied. From experimental data it was concluded that: (1) Solid phase welding accounts for almost the entire magnitude of the static friction coefficient for metal-metal couples. (2) Fe-Ag does not solid phase weld, other insoluble couples examined do weld. (3) A measure of solid phase weldability appears to be the quotient of the work of adhesion and the shear strength of the weaker couple. (4) Titanium does not solid phase weld to crystalline oxides to any greater extent than copper. (5) Freshly cut titanium solid phase welds to other metals in frictional contact. The degree of welding is independent of the atmosphere (air, nitrogen, or argon) or the time (less than 24 hours) at room temperature in such atmospheres. This behavior is contrary to that experienced with iron or copper. With the latter couples less than five minutes exposure of a freshly cut surface to air is sufficient to markedly reduce the degree of solid phase welding. (6) The seizability of titanium is probably due to the lack of resistance, of the contaminant film formed on it in air, to penetration or wear in frictional contact.

INTRODUCTION

THIS project has as its general objective a determination of the factors that affect seizure and galling. As a first step toward the fulfillment of this objective, it was decided to study the solid phase weldability of different metals to determine whether there are exceptions to the rule that all clean metals weld to each other. The literature reveals many conflicting statements making a conclusion on this question worthwhile. Specifically, Bowden and Tabor imply that all clean metals solid phase weld to each other.[1] Merchant[2] disputes this view, claiming that metals which are mutually insoluble do not weld to each other. Feng[3] believes that solid phase welding is unimportant in friction, the major contribution to the friction force arising from mechanical interlocking of the two bodies in contact.

The experiments to be described have been designed to provide a conclusive answer to this controversy. Also, because of the current interest in titanium, it has been included in the list of metals to be tested.

EXPERIMENTAL TECHNIQUES

Friction Apparatus

A modified Bowden-Leben apparatus for measuring friction coefficients has been constructed and is illustrated in Figs. 1(a) and (b). The shaper A is used to cut a clean surface on hard slider specimen B attached to piston C of hydraulic cylinder D. The box E maintains an inert atmosphere supplied from a tank of prepurified nitrogen or argon through a titanium chip furnace train to remove oxygen. After the shaper cuts a clean

[1] F. P. Bowden and D. Tabor, "Wear and Damage of Metal Surfaces," *Mechanical Wear*, Editor, J. T. Burwell, Jr. (American Society for Metals, New York, 1950), p. 117.
[2] M. E. Merchant, Discussion to reference 1, p. 135.
[3] I. Ming-Feng, J. Appl. Phys. 23, 1011 (1952).

surface on the slider, the piston C is moved forward under hydraulic pressure at a rate controlled by bleeding oil through a needle valve. The hydraulic pressure is supplied by a tank of nitrogen acting through the medium of oil contained in a reservoir. When the slider is properly under the hemispherical rider F, the motion of the slider is stopped. The hemispherical rider F is then lowered by means of screw G into contact with the slider. The load is applied by deflection of member H and measured by means of strain gauges and the Sanborn recording bridge I. When the desired load is reached clamp J is tightened and the slider again set into motion by means of the hydraulic action previously described. The frictional force is then measured by means of the deflexion of flexure member K, attached strain gauges and another channel on the Sanborn recorder.

Friction Tests—Materials and Methods

The materials used in these tests are high-purity elements, with the exception of the copper and titanium used for the sliders. These elements and their nominal purities, as received, are listed in Table I. Some of these elements were in powder or shot form and had to be made into usable shape. Melting and solidification into the desired shape was accomplished in cleaned, evacuated Pyrex tubes. Fine grained specimens were obtained from this treatment. The Brinell hardnesses of the specimens, as used, are also listed in Table I.

The four hardest materials, Ti, Fe, Ag, and Cu are used as flat slider materials and are cut in place by the shaper to provide clean surfaces. The other materials and also Ag and high-purity annealed Cu are used to make the softer hemispherical rider. The riders because of their relative softness are self-cleaning. The harder slider acts to cut or wear away the surface of the mating softer rider during a friction test. It is important to note however, that the rider is never run twice over the same path on the slider. Thus, if no material transfer takes place from the hard slider to the soft rider the type of bond is always a rider-slider bond and never a rider-rider or slider-slider bond.

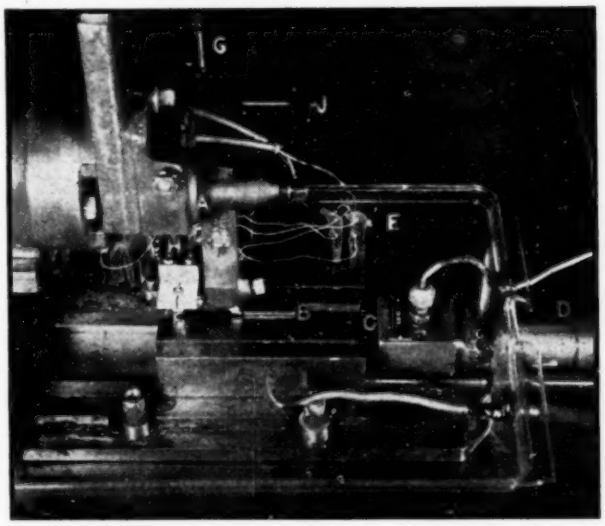


FIG. 1(a). Friction measuring apparatus (close up view). *A*. Shaper. *B*. Slider. *C*. Piston. *D*. Hydraulic cylinder. *E*. Atmosphere box. F. Hemispherical slider. *G*. Load adjustment screw. *H*. Load flexible member. *J*. Load clamp.

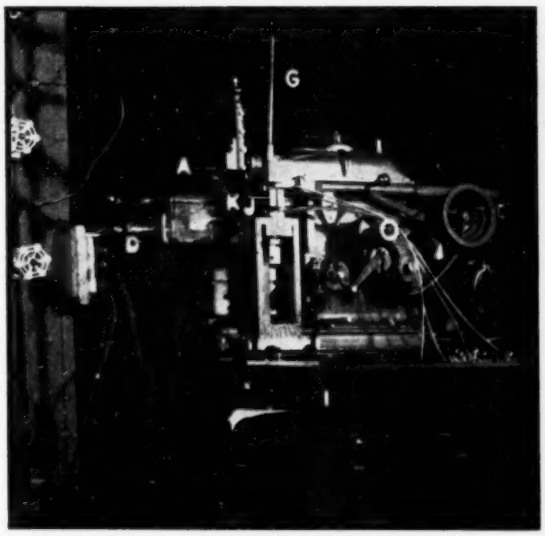


FIG. 1(b). Friction apparatus. *A*. Shaper. *D*. Hydraulic cylinder. *E*. Atmosphere box. *G*. Load adjustment screw. *I*. Sanborn recorder. *J*. Load clamp. *K*. Friction flexible member (enclosed within rubber gasket).

Friction Tests—Procedure

1. A clean surface was cut on the slider in air.
2. At least 15 minutes elapsed before making friction runs on this surface.
3. At least 3 friction runs were made in air.
4. The box surrounding the friction equipment was then partially evacuated and either prepurified, deoxidized nitrogen or argon was allowed to flush through the box for at least 20 minutes.

TABLE I.

Metal	Purity	Brinell hardness
Lead	99.999	3.2
Copper rider (annealed)	99.999	9
Zinc	99.999	12
Magnesium	99.999	17
Cadmium	99.999	21
Aluminum	99.99	27
Iron	99.999	57
Copper rider (unannealed)	99.999	57
Copper slider (Electrolytic Tough Pitch)	99.9+	62
Silver	99.9	70
Titanium	99.+	189

TABLE II. Static coefficients of friction.

Friction components Rider on slider	Data obtained in air: 1st run	Data obtained in air: Succeeding runs	Data obtained in N_2 or argon: Slider clean	Data obtained in N_2 or argon: Rider + slider clean	Data obtained in air on surfaces produced in nitrogen or argon
Al on Ti (3)[a]	0.52–0.65	0.67	0.68	0.69–0.87	0.65
Al on Fe (3)	0.41	0.44	0.59	0.60	0.45
Al on Cu (1)	...	...	0.54	0.90	0.45
Ag on Ti (3)	0.48±0.1	0.51	0.60	0.66	0.54±0.1
Ag on Fe (2)	0.41±0.1	0.44	0.46	0.48	0.47
Ag on Cu (1)	...	...	0.29[b]	0.75	0.29
Cd on Ti (2)	0.64	0.69	0.68	0.75	0.71
Cd on Fe (3)	0.57	0.55±0.1	0.57±0.1	0.71	0.59
Cd on Cu (1)	...	...	0.74	0.86	0.73
Cu on Ti (2)	0.62[b]	0.62[b]	0.66[b]	0.65	0.62[b]
Cu on Fe (3)	0.51±0.1	0.57±0.1	0.73	0.72	0.64
Cu on Cu (2)	0.45±0.1	0.59	0.67±0.1	1.65±0.5	1.30–0.7[c]
Mg on Ti (3)	0.71	0.67	0.67	0.68	0.66
Mg on Fe (2)	0.42	0.47	0.44	0.45	0.44
Mg on Cu (3)	...	...	0.35	0.40	0.47
Pb on Ti (4)	0.63±0.1	0.65±0.1	0.68	0.70	0.67
Pb on Fe (4)	0.76–1.44	0.59–0.96	0.65–0.94	0.77–0.95	0.51–1.16
Pb on Cu (2)	0.42[b]	0.55[b]	0.55[b]	1.28[b]	0.85[b]
Zn on Ti (2)	0.58	0.65	0.67	0.73	0.67
Zn on Fe (2)	0.54	0.57	0.65	0.66	0.62
Zn on Cu (1)	...	...	0.69	0.90	0.54
Ti on Ti (1)	0.69	0.65	0.68	0.67	0.68
Fe on Fe (3)	0.17–0.51	0.51	0.62	0.69	0.53
Ag on Ag (2)	...	...	1.76	2.44	1.16

Slider cut here in air | Slider cut here in nitrogen | Rider rubbed here on file in nitrogen | Air admitted here to box

The general uncertainty in each value is ±0.05 except where noted differently.
[a] Number of separate runs to determine average values shown in table.
[b] Stick-slip absent.
[c] Value decreases with increased holding time in atmosphere.

5. A friction run was then made on the same surface now in an inert atmosphere. If the friction value was the same as obtained in air then a new surface was cut on the slider, the cut being made in the inert atmosphere. (In all cases no change in the friction coefficient using the oxidized slider was found as a result of changing the atmosphere.)

6. At least 1 friction run was made on this freshly cut surface.

7. The rider was then rubbed on a clean fine file at high load and speed to clean the rider and at least 2 friction runs were carried out for this condition.

8. The box was opened up to air, 5–15 minutes allowed to elapse and another friction run made to check the results obtained in item 3.

9. The friction tracks were then examined using a microscope.

Friction Tests—Data

A summary of the friction data obtained with the above described procedure is presented in Table II for loads averaging between 30 and 120 grams and relative sliding speeds varying between 0.5–1 cm per minute. Variations of speed and load within this range were found to have little significant effect on the friction data. Stick-slip was found in most of the couples studied, except as noted. The friction values in the table correspond to the maximum values obtained during the stick portion of the stick-slip cycle.

Results have also been obtained at relatively high loads with much the same results, except implications of surface melting for the cases of Pb, Cd, and Zn riders on Ti. The results for the lower loads are probably representative of the solid surfaces.

The effect of the hardness of the rider for a given rider material was also investigated for Cu on Ti. The cold worked copper having a hardness of 57 Brinell was annealed by heating to 550°C for 25 minutes to a hardness of 9 Brinell. The friction data for this series of experiments is presented in Table III.

The reproducibility of the friction data was adequate. All data were within ±0.05 of the values shown in Table II except where noted.

An aid to the interpretation of Table II is presented in Table IV where the increment in the static friction coefficient on provision of a freshly cut slider surface* (Δ_1) and of a freshly cut rider surface (Δ_2) are shown.

The results presented in Tables II–IV may be described as follows:

1. An increase in the friction coefficient on provision of a "clean" Fe or Cu surface in an inert atmosphere was observed for the rider metals, Al, Cu, and Zn.

2. A similar increase on provision of a "clean" Ti surface was observed only for the Ag rider.

3. An increase in the friction coefficient on provision of a "clean" rider surface on a "clean" slider surface was observed for the case of Al, Ag, Cd, and Zn riders on Ti and Cu sliders and for the case of a Cd rider on an Fe slider.

* The provision of a freshly cut slider surface and a freshly cut rider surface, in an inert atmosphere, shall be referred to as cleaning of the slider and rider metals, respectively.

TABLE III. Annealed *versus* unannealed Cu on Ti.

	Coefficient of static friction				Average difference between static and kinetic coefficients			
	Cut in air		Cut in inert atmosphere		Cut in air		Cut in inert atmosphere	
	1st run in air	Succeeding runs	Run in I.A.	Runs after rubbing rider	1st run in air	Succeeding runs	Run in I.A.	Runs after rubbing rider
Unannealed	0.50	0.56	0.51	0.63	0.07	0.00	0.06	0.04
	0.50	0.55		0.69	0.00	0.00		0.06
(57 Brinell)	0.57	0.53	0.50	0.50	0.19	0.03	0.00	0.02
	0.51	0.49	0.47	0.60	0.00	0.00	0.02	0.04
Annealed	0.46	0.55	0.58	0.66	0.06	0.05	0.05	0.05
	0.49	0.48	0.46	0.54	0.03	0.03	0.02	0.05
(9 Brinell)	0.57	0.52	0.52	0.57	0.03	0.02	0.04	0.03

Load —1 kg
Speed—0.5–1 cm per minute

4. With both rider and slider "clean" the friction coefficients for a given rider metal were about the same on both Ti and Fe sliders for Cd, Cu, Pb, and Zn. (Al probably falls into this category as well.)

5. Only for the case of Ag and Mg riders were the values for the "clean" state higher on Ti than on Fe.

6. Ti run on itself showed a constant value of the friction coefficient whether "clean" or "unclean."

7. Fe or Cu run on itself showed an appreciably higher friction coefficient in the "clean" state as compared to the "unclean" state.

8. Values of the friction coefficient for the Cu slider in the clean state are higher than for the same rider metals on Ti and Fe sliders.

GUIDE TO THE INTERPRETATION OF FRICTION COEFFICIENTS

Bowden has shown that if welding is assumed to occur between the friction components then the static friction coefficient is related to the shear strength and the flow pressure (hardness) by the equation

$$\mu = \alpha \frac{S}{H}, \qquad (1)$$

where μ static friction coefficient, S shear strength of weaker component, H flow pressure of softer component, α fraction of contact area which has welded.

According to Eq. (1) for the same softer and weaker rider metal sliding over different harder metals, the coefficient of friction should be independent of the slider metal provided that α, the fraction of the contact area that is clean (welded) does not change. For what has been called the clean state in this paper it has been found that the friction coefficient is the same for Fe and Ti sliders, but is generally higher for Cu sliders, with given rider metals. This result can be interpreted as indicating that α in the clean state for Fe and Ti is the same, but is larger for Cu. This interpretation is consistent with the fact that after thorough cleaning in high-vacuum friction coefficients can approach the magnitude of 100. Thus, variation in cleanliness alters the factor α, as indicated by the finite values of Δ_1 and Δ_2 in Table IV.

ANALYSIS OF RESULTS

Correlation between Friction Data and Work of Adhesion

Some of the questions raised in the introduction can be conclusively answered on the basis of the results obtained in this investigation. There is no doubt that solid phase welding occurs in frictional contact. The mere contact of a clean surface with air at room temperature markedly reduces the friction coefficient in many cases. The amount of such reduction is a function of the time the surface is bathed by the air. This reduction is accomplished without any observable change in surface geometry. Increasing the temperature at which the surface is held in the air without any visible oxide film being formed and without any observable change in surface geometry reduces the friction coefficient from 0.7 to about 0.2 (see Table V). Thus, the assumption of Feng is found to be false because mechanical interlocking cannot explain the above results and thus can play only a minor role in friction phenomena. The friction couple used in this study, a soft hemispherical rider on a hard flat slider, should magnify this effect if present. The conclusion is therefore applicable to other types of friction couples.

Also, the assertion of Merchant, that solid phase welding will not occur between metals that are mutually insoluble, has been found inaccurate. Cd which is insoluble in Fe, welds to Fe as well as any other metal tested (for example, Zn and Cu, both of which are

TABLE IV. Effect of cleaning surfaces in nitrogen on static friction coefficients.

	Ti		Fe		Cu	
	Δ_1[a]	Δ_2[b]	Δ_1	Δ_2	Δ_1	Δ_2
Al	0.01	0.20	0.15	0.01	0.09	0.36
Ag	0.09	0.06	0.02	0.02	0.00	0.46
Cd	0	0.05	0	0.14	0.01	0.12
Cu	0.04	0	0.16	0	...	1.00
Mg	0	0	0	0	0	0
Pb	0.02	0.02	...	...	0.00	0.73
Zn	0.02	0.06	0.08	0.01	0.15	0.21

[a] Increase in static friction coefficient on cleaning slider in nitrogen.
[b] Increase in static friction coefficient on cleaning rider and slider in nitrogen.

TABLE V. Effect of heating Ti slider in air on static friction coefficient.

Friction components	Unheated (in air)	Heated[a] (in air)
Cu on Ti	0.65	0.23
Mg on Ti	0.68	0.23

[a] Ti surface heated in air at 350°C for 17 hours, rider surface was cleaned by rubbing prior to test. No change in coefficients observed on performing test in air or in argon.

soluble in Fe.) Some insoluble metals however, do not weld to each other. In particular, the rider metal Ag does not weld to Fe. The latter result also contradicts Bowden and Tabor's implication that all metals in frictional contact solid phase weld to each other.

In the terminology used in this investigation solid-phase welding takes place if separation of the two friction components does not occur at the interface but rather within the weaker component. Thus, to predict whether a given friction couple will solid phase weld or not it is necessary to know the relative strengths of the weaker component and the interphase interface, under the conditions of stress imposed on the friction couple. These stress conditions are not well defined and hence it is difficult to state precisely the condition of antisolid phase welding. From the following argument those friction pairs which will certainly solid phase weld can be predicted.

It is known that estimates of strength using the work necessary to form two surfaces in place of an interface,[4] the work of cohesion, are larger than observed strengths by a factor of at least 100. Thus, it can be assumed that if the work required to separate the friction couple at the interphase interface is as large as the ideal work required to separate the weaker component at a grain boundary, then separation will certainly occur within the weaker component. This assertion is based on the assumption that there are no imperfections at an interphase interface, which can lead to a lower strength than that associated with the imperfections in a grain boundary, or the interior of the weaker component. The work of separation at the interphase interface is the work of adhesion in surface chemistry nomenclature, i.e.,

$$W_{\alpha\beta}=\gamma_\alpha+\gamma_\beta-\gamma_{\alpha\beta},$$

where γ_i surface free energy of phase i per unit area, $\gamma_{\alpha\beta}$ interphase interface free energy per unit area. Thus, if $W_{\alpha\beta}\geq W_{\alpha\alpha}$, solid phase welding will occur. (In this case, α is assumed to be the weaker phase.) Bondi has collected values of $W_{\alpha\beta}$ and $W_{\alpha\alpha}$.[5] These values are shown in Table VI together with the conclusions drawn from friction measurements concerning solid phase weldability. It is apparent that solid phase welding occurs for the couples which have work of adhesion values on the order of work of cohesion values. Also, it is apparent that couples, which have work of adhesion values less than about $\frac{1}{3}$ the corresponding work of cohesion, solid phase weld to a lesser extent, if at all. In this respect it is interesting to note that metal-oxide interfaces, which are normally considered to be resistant to solid phase welding, have corresponding work of adhesion values on the order of 500 ergs/cm² or less. This value is about the same magnitude found for the Ag-Fe couple. The couples which weld have work of adhesion values exceeding about 1500 ergs/cm². This correlation is not strictly rigorous because the strength properties of the weaker component will govern the critical value below which welding will not occur and above which welding will occur for the metal in question. Perhaps a more representative criterion would be the quotient of the work of adhesion and the strength of the weaker component.

[4] F. Seitz, *Physics of Metals* (McGraw-Hill Book Company, Inc., New York, 1943), p. 167.

[5] A. Bondi, Chem. Revs. 52, 417 (1953).

Seizability of Titanium

According to the results shown in Tables II–IV the cleanliness of the Ti surface is not changed by cutting in air, cutting in argon or nitrogen, or subsequent exposure of the surfaces prepared in the latter condition to air. Values of Δ_1 for the Ti slider are nil except for the Ag rider. The Fe surface however, changes its cleanliness as shown by the finite values of the quantity Δ_1 for the Fe slider shown in Table IV. Similarly for the Cu surface, the question arises as to whether the Ti surface is equally "unclean" or equally "clean" in the atmospheres. The friction data for the Cd, Cu, Pb, and Zn riders in the "clean" condition, which have the same values on both the Fe and Ti surfaces imply that welding is occurring equally well to the Ti surfaces as to the Fe surfaces, even in air. Thus, it appears that the Ti surfaces are effectively clean in that whatever layer exists on the Ti surface does not prevent solid phase welding to the degree that the layer produced on Fe in air prevents such welding. In order to check this idea a surface was machined on a Ti specimen in air according to the same procedure used for the data in Table II. This specimen was then heated in air at 350°C for 17 hours. No distinct oxide layer was observed visually, however, this treatment so altered the Ti surface as to

TABLE VI. Correlation between work of adhesion and solid phase weldability.

Friction components	Static friction coefficients: Both surfaces clean	Same surfaces as tested in air	Work of adhesion ergs/cm²
Ag on Cu	0.75	0.29	2330
Cu on Fe	0.72	0.50	2675
Ag on Fe	0.48	0.47	712
Cu on SiO_2	0.10	0.10	325

Work of cohesion at grain boundaries, ergs/cm²	
Metal	
Ag	1780
Cu	2260

reduce the friction coefficients for the Cu-Ti and Mg-Ti couples from about 0.7 to about 0.2 as shown in Table V. Thus, it can be concluded that a freshly machined surface of Ti is clean in the sense that no "lubricating" oxide layer exists on it. Such a layer can be produced by heating the Ti in air.

A question still remains to be answered. Although no lubricating quality oxide film may exist on the Ti surface, such oxides may exist on the rider surfaces, in air. However, the coefficients measured in air (after the very first run) are the same as those for clean surfaces for many of the couples comprising Ti. The question is whether Ti can weld to some oxides as well as it can to the clean metal. In order to check this point, Ti and Cu riders were run against a number of crystalline metal oxides. The data are presented in Table VII. It is apparent that the friction coefficients are the same, within the experimental uncertainty, for both the Cu and Ti riders, with the possible exception of Ti on rutile. From this result, it must be concluded that Ti does not weld to oxides to any greater extent than other metals, and, if welding occurs, the degree of welding is much less than it is for Ti-metal couples.† The abnormal seizability of titanium must therefore be ascribed to the thinness (or absence) of the contaminant (oxide?) film formed on it in air, at room temperature. This film, if it exists, cannot be resistant to penetration or removal under sliding contact. This conclusion is reasonable in view of the fact that the solubility of oxygen in titanium is extremely large. Thus, the formation of an oxide film at room temperature would require the oxygen concentration at the surface to reach the value at the terminal phase boundary—a process which may require considerable time at room temperature.

TABLE VII. Values of static friction coefficients against crystalline oxides.

	Tested in air	Tested in nitrogen Rider unclean	Rider clean
Ti on corundum (Al_2O_3)	0.40–0.50	0.40	0.40–0.60
Ti on quartz (SiO_2)	0.15[a]	0.15[a]	0.15[a]
Ti on zircon (ZrO_2)	0.30	0.30	0.30
Ti on rutile (TiO_2)	0.20	0.25	0.40
Ti on magnetite (Fe_3O_4)	0.20	0.25	0.25
Ti on hematite (Fe_2O_3)	0.15[a]	0.15[a]	0.15[a]
Cu on corundum	0.55	0.35	0.45
Cu on quartz	0.10[a]	0.15[a]	0.10[a]
Cu on zircon	0.20	0.25	0.25
Cu on rutile	0.20	0.20	0.20
Cu on magnetite	0.25	0.25	0.25
Al on magnetite	0.20	0.20	0.20
Zn on magnetite	0.20	0.25	0.30
Cu on hematite	0.20[a]	0.20[a]	0.25[a]

[a] Stick-slip absent.

Summary of Results and Conclusions

1. Solid phase welding accounts for almost the entire magnitude of the static friction coefficient for most metal-metal couples.

2. Some couples, comprised of metals which are mutually insoluble in the solid state, solid phase weld; some do not solid phase weld.

3. A measure of the solid phase weldability appears to be the quotient of the work of adhesion to the strength of the weaker component.

4. Titanium does not solid phase weld to oxides to any greater extent than Cu.

5. Freshly cut titanium solid phase welds to other metals in frictional contact. The degree of welding is independent of the atmosphere (air, nitrogen, or argon) or the time (less than 24 hours) at room temperature in such atmospheres. This behavior is contrary to that experienced with iron or copper. With the latter couples less than five minutes exposure of a freshly cut surface to air is sufficient to markedly reduce the degree of solid phase welding.

6. The seizability of titanium is probably due to the lack of resistance, of the contaminant film formed on it in air, to penetration or wear in frictional contact.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of the Wright Air Development Center, A.F., Department of Defense, through Contract AF-18(600)-75, and their indebtedness to Robert Duncan and James Kauzlarich for help in various phases of this work. We would also like to thank the Geology Department of Columbia University for loan of the metal oxide crystals used in these experiments.

† This conclusion may need to be modified for the couple Ti-MgO. Table II shows that for the couples Mg-Fe and Mg-Cu the friction coefficients are low and unchanged by cleaning. This result is interpreted to mean that the Mg picks up oxygen even from the inert atmosphere quickly enough to form an effective protective film to Fe and Cu, but because of the high friction coefficient on Ti, this film is not protective against Ti.

On the Band Width of Cavity Antennas*

MARSHALL H. COHEN
Ohio State University, Columbus, Ohio
(Received August 10, 1953)

The objective of this work is to determine the orders of magnitude of the band width available from a cavity antenna. For simplicity of calculation a piece of shorted square wave guide of side a, opening into a half-space, is considered, and the resonance frequency and band width $(1/Q)$ are calculated from transmission-line theory. The analysis depends on an equivalent aperture admittance which is defined for the lowest-order wave guide mode.

This type of antenna has a minimum Q which is equal to $0.424(a/\lambda)^{-3}$ when $(a/\lambda)<0.35$. The cavity depth required for the minimum Q is of the order of $0.4a$. The required dielectric constant increases rapidly as a/λ decreases. The effect of dielectric loss is considered. Some experimental results which agree well with theory are presented.

INTRODUCTION

IT is well known that the band width of an electrically small cavity antenna is severely limited by the aperture dimensions. However, many theoretical aspects of this phenomenon are undeveloped. In the present investigation the resonance frequency and band width will be calculated for a length of shorted square wave guide opening into a half-space. This is a restricted class of antennas, but the results are indicative of what can be expected from other practical shapes. It will be shown that this configuration has a maximum band width which increases as the cube of the frequency.

Most previous work on this subject has used the stored energy definition of Q. Chu[1] used the orthogonality of spherical waves to set up an equivalent circuit from which he found relations among size, gain, and Q. Counter[2] considered antennas of the type considered here, and attempted to find the stored energy by a direct integration procedure. Some of his results are similar to those reported here.

In this paper Q is found from the normal modes. Such modes are the possible natural oscillations which are more commonly associated with closed systems such as networks or resonant cavities, but which nevertheless are assumed to exist for antennas. Heretofore, they have been discussed only for the simplest types of structures, including spheres, spheroids, and thin wires.[3,4]

It is assumed that an antenna has an infinite discrete set of modes and that their time dependence is of the form $\exp(jp_n t)$, where p_n is a complex number,

$$p_n=\omega_n+j\xi_n, \quad \xi_n>0. \tag{1}$$

Each mode has a Q defined by

$$Q_n=\omega_n/(2\xi_n). \tag{2}$$

The complex frequency p_n allows us to keep the formal solutions to the wave equation, but some new interpretations are necessary. For example, it can be shown that a normal mode wave has an exponential envelope which increases in the direction of propagation. The mathematical expression describes the wave as increasing without limit, but since the corresponding physical situation (transient oscillation) was started at a finite time in the past, the wave will be zero outside of some sphere surrounding the antenna. Further interesting effects occur when the cavity is a section of uniform wave guide.[5]

The customary proofs of orthogonality and completeness cannot be applied to the set of normal modes, because of the behavior at infinity. Hence it cannot be said that any transient oscillation (including the steady-state condition) is representable as a linear combination of the modes. However, the usual procedure in resonance problems is to assume that each steady-state resonance corresponds to one of the normal modes. We will do this with cavity antennas.

CALCULATIONS FOR A SQUARE WAVE GUIDE CAVITY ANTENNA

Consider a piece of square wave guide, shorted at one end and opening into a half-space at the other (Fig. 1). This antenna may be treated on a transmission line basis, with terminals at the aperture, provided the field in the wave guide consists essentially of one wave-guide mode, and the aperture admittance is known for that mode.

In the following we assume that these conditions are met. Therefore, since there are no sources, the system (Fig. 1) oscillates when

$$Y_n+y_n\coth(\gamma_n l)=0, \tag{3}$$

where n=antenna (normal) mode index, Y_n=aperture

* This paper is a portion of the dissertation presented by the author as part of the requirements for the degree Doctor of Philosophy at The Ohio State University. The dissertation has been published as a technical report of the Antenna Laboratory (see reference 5).

[1] L. J. Chu, J. Appl. Phys. **19**, 1163 (1948).

[2] V. A. Counter, Miniature Cavity Antennas, Quarterly Reports Nos. 2–6, Microwave Laboratory, Stanford University, 1948–1950.

[3] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941), p. 554.

[4] S. A. Schelkunoff, *Advanced Antenna Theory* (John Wiley and Sons, Inc., New York, 1952), Chap. 6.

[5] M. H. Cohen, The Normal Modes of Cavity Antennas, Technical Report 486–7, Antenna Laboratory, Ohio State University Research Foundation, Columbus, Ohio, December, 1952.

admittance, y_n = wave-guide characteristic admittance, γ_n = wave-guide propagation constant. The wave-guide parameters are given by[6]

$$\begin{aligned} &y_n=\gamma_n/(jp_n\mu) \text{ for } TE \text{ modes},\\ &y_n=jp_n\epsilon/\gamma_n \text{ for } TM \text{ modes},\\ &\gamma_n^2=\kappa^2-k_n^2, \end{aligned} \tag{4}$$

where $k_n^2=p_n^2\mu\epsilon$, $\kappa=(2\pi)\times$cut-off wave number of wave-guide mode.

Equation (3) may also be obtained by a variational calculation for the eigenvalues p_n, under the assumption that the distribution of tangential electric field in the aperture is a single wave-guide mode function.[5] A similar variational calculation[7] for the aperture admittance for a square wave guide, under the assumption that the tangential aperture field is the TE_{10} mode distribution, yields the following for the aperture conductance G and susceptance B:

$$\begin{aligned} G=&0.00901x^2-0.01522x^4+0.01436x^6\\ &-0.00848x^8+\cdots\\ B=&-0.000809x^{-1}+0.00537x-0.01278x^3\\ &+0.01596x^5-0.01161x^7+\cdots, \end{aligned} \tag{5}$$

where

$$\begin{aligned} &x=ap/(2\pi c),\\ &c=\text{velocity of light in vacuum}. \end{aligned} \tag{6}$$

When p is real, $x=a/\lambda$.

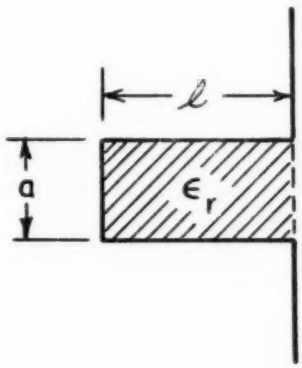


FIG. 1. Square wave-guide cavity antenna.

Let the medium in the wave guide have the electrical constants $\mu=\mu_0$ and $\epsilon=\epsilon_r\epsilon_0$, where ϵ_r is real. Then, in terms of the dimensionless variables x, l/a, and ϵ_r, the eigenvalue equation [Eq. (3)] for the TE_{10} mode becomes

$$f(x)=-j(\epsilon_0/\mu_0)^{\frac{1}{2}}(2\pi x)^{-1}\beta a\cot\beta l+G(x)+jB(x)=0, \tag{7}$$

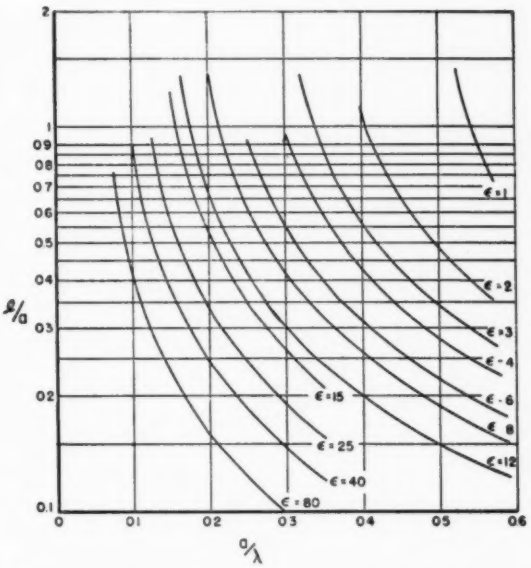


FIG. 2. Relations among cavity parameters for first normal mode of square wave-guide cavity antenna.

where

$$\beta=-j\gamma=(\pi/a)(4\epsilon_r x^2-1)^{\frac{1}{2}}. \tag{8}$$

A solution x_n of Eq. (7) is obtained by expanding $f(x)$ in a Taylor's Series about the value x_{0n} which is obtained when $G=0$. For the range of values under consideration the first two terms of the series give adequate accuracy.[8] Thus x_n is given by

$$x_n=x_{0n}+\frac{-G(x_{0n})}{\partial f/\partial x_{0n}}. \tag{9}$$

The real part of x_n is a/λ_n, and the imaginary part gives the decay constant. The corresponding Q_n is

$$Q_n=\frac{\omega_n}{2\xi_n}=\frac{a/\lambda_n}{2Im(x_n)}. \tag{10}$$

The calculations for the first mode indicate that when Q is high ($Q>7$) the following approximations are valid:

$$a/\lambda_n\approx a/\lambda_{0n} \tag{11}$$

$$Q_n\approx\frac{\omega_n}{2G}\frac{\partial B_T}{\partial\omega_n}, \tag{12}$$

where B_T is the total admittance at the aperture. It is interesting to observe that Eq. (12) is the same as the approximate equation for Q obtained from circuit theory.

[6] N. Marcuvitz, *Waveguide Handbook* (McGraw-Hill Book Company, Inc., New York, 1951), M.I.T. Radiation Laboratory Series, Vol. 10, Chap. 1.

[7] Cohen, Crowley, and Levis, The Aperture Admittance of a Rectangular Waveguide Radiating into Half-Space, Project Report 339–22, Antenna Laboratory, The Ohio State University Research Foundation, Columbus, Ohio, November 1951. For an alternative treatment, see L. Lewin, *Advanced Theory of Waveguides* (Iliffe and Sons, Ltd., London, 1951), p. 121.

[8] The validity of keeping only the linear terms was checked by computing the second derivative at one point. The point chosen was $a/\lambda=0.60$, $\epsilon_r=1$; this is a point of very low Q and the approximation should be worst here. The inclusion of the second derivative lowered Q by 3 percent and lowered a/λ by 0.7 percent.

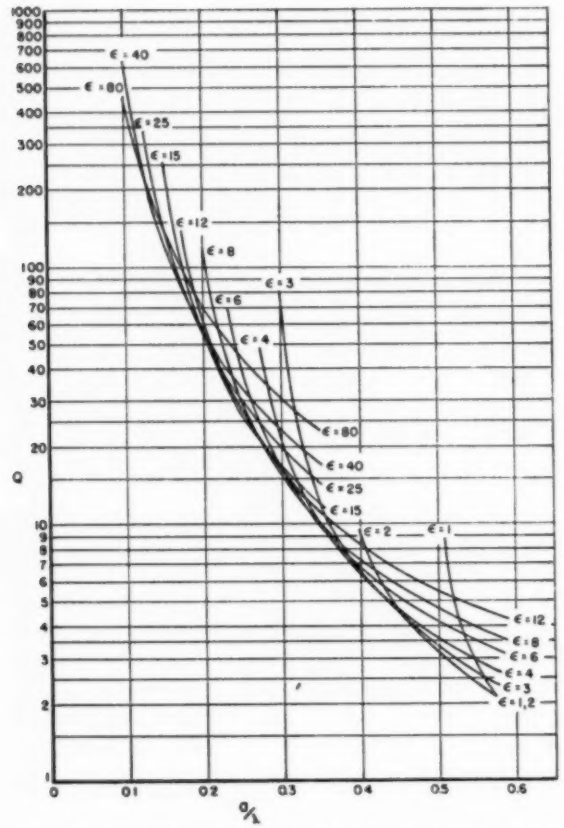


FIG. 3. Resonance characteristics of first normal mode of square wave-guide cavity antenna.

Results for First Normal Mode

A series of numerical calculations was made for the first normal mode of the square wave-guide cavity antenna, according to the above procedure. The parameters of interest are l/a, ϵ_r, a/λ, and Q. (The subscript n has been dropped for simplicity.) Since it is often desired to design a cavity for a specific resonance frequency, a/λ is chosen as an independent variable.

Figure 2 shows the dependence of l/a on a/λ for various values of ϵ_r. It is seen that there is a variety of pairs of ϵ_r and l/a which produce the same a/λ. The individual curves asymptotically approach that value of a/λ which corresponds to cutoff.

The values of Q which result from the various configurations of Fig. 2 are shown in Fig. 3. The envelope of the family of curves in this graph corresponds to the minimum Q. Each value of a/λ has a specific pair of values of ϵ_r and l/a which are necessary to obtain the minimum Q.

The minimum Q was separately evaluated by maximizing the denominator in Eq. (10) with respect to ϵ_r and l/a. The result is shown in Fig. 4, where the minimum Q is represented by the solid line. It is interesting to note that this is closely approximated by the formula

$$Q_{\min}=0.424(a/\lambda)^{-3}, \qquad (13)$$

which is represented by the dashed line in Fig. 4. The cubic dependence may be readily derived by considering the aperture admittance to be small perturbation from an open circuit.[9] The absolute values of $Q_{\min}$ so obtained, however, are somewhat smaller than those given by (13).

Figures 5 and 6 give the values of l/a and ϵ_r which produce the minimum Q. Note that these values are not critical, since Q is a relatively slow function of ϵ_r, near the minimum Q.

Some typical numerical results for the first mode are shown in Table I. The first four columns include results which are represented in the graphs. The fifth column, βl, is the electrical length of the cavity at resonance. The columns labeled stored energy are obtained by assuming that $\partial B_T/\partial\omega$ is proportional to the total stored energy of the system, and that it may be split in the same way as the susceptance, i.e., into external and internal portions. This assumption is, of course, an approximation, which is valid for high-Q antennas. It becomes true as Q increases and the frequency becomes real. It is seen from Table I that most of the stored energy is in the cavity. This was noticed in every calculation that was made for Q, including those for higher modes.

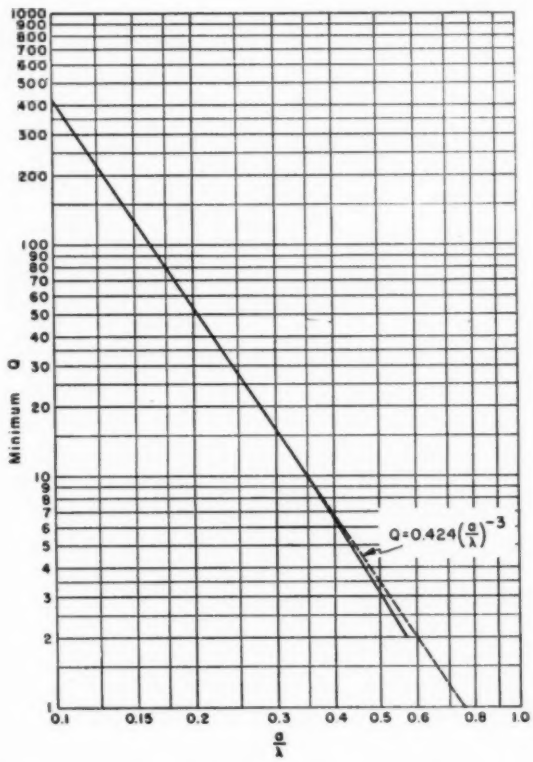


FIG. 4. Minimum Q *versus* a/λ for first normal mode of square wave-guide cavity antenna.

[9] Counter (reference 2) essentially used an open circuit boundary condition. He also obtained a cubic dependence for minimum Q.

Effect of Dielectric Loss

Losses in the dielectric in effect simply add a term to the aperture conductance. When Q is high in place of (12) we now have

$$Q_n \approx \frac{\omega_n}{2(G+\mathcal{G})} \frac{\partial B_T}{\partial \omega_n}. \tag{14}$$

The dielectric loss does not effect the resonance frequency so long as Q remains high, but it may have an appreciable effect on Q itself. The additional term $\mathcal{G}$ may be found by letting ϵ_r be a complex number and expanding $f(x,\epsilon_r)$ about the points x_{0n} which were used previously.

From Eq. (14) one may write

$$1/Q_T = 1/Q_R + 1/Q_D, \tag{15}$$

where Q_T is the total Q, Q_R is the radiation Q given by Eq. (12), and Q_D is the dielectric Q. When Q_T is high

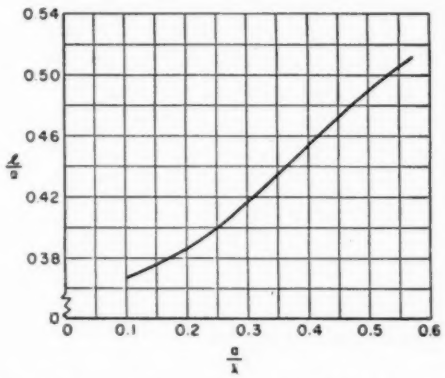


FIG. 5. Geometry necessary to obtain minimum Q.

it may be shown that

$$1/Q_D \approx \tan\delta, \tag{16}$$

where

$$\tan\delta = \text{loss factor} = \sigma/(\omega\epsilon). \tag{17}$$

Hence we have the approximate formula

$$1/Q_T \approx 1/Q_R + \tan\delta. \tag{18}$$

The radiation efficiency is Q_T/Q_R; hence

$$\text{efficiency} \approx (1+Q_R \tan\delta)^{-1}. \tag{19}$$

This formula illustrates the principle that a high-Q system fundamentally has a low efficiency.

CORRELATION WITH STEADY-STATE BEHAVIOR

The problem now arises as to what these modes have to do with the steady-state antenna. An extensive literature on this subject already exists,[10–15] and it has been

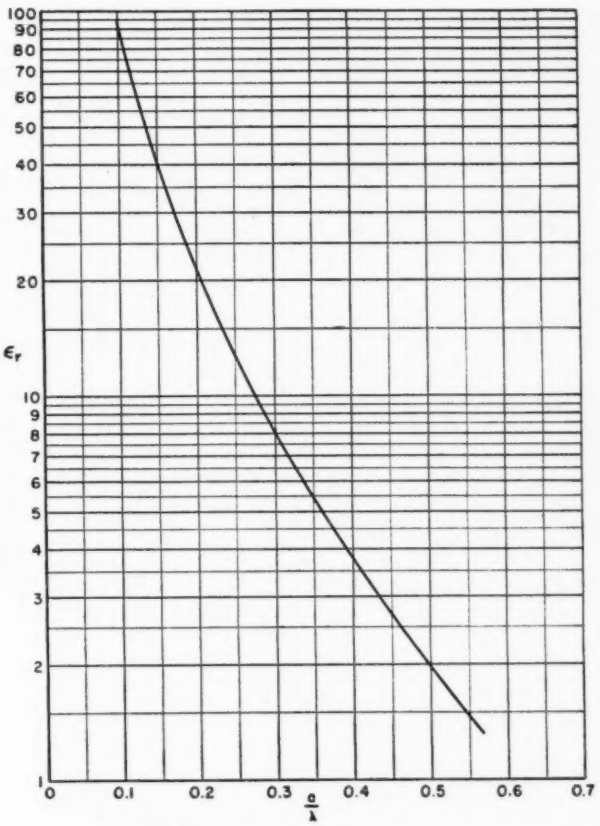


FIG. 6. Dielectric constant to obtain minimum Q.

shown that when the antenna has a high Q it is approximately represented by the network in Fig. 7. The RLC circuit is supposed to represent the cavity in both the steady-state and transient cases, and thereby the connection between normal modes and impedance of the antenna is made. Thus, the normal mode oscillation frequency and Q are equal to the steady-state resonance

TABLE I. Typical results for first normal mode of square cavity antenna.

a/λ	ϵ_r	l/a	Q	βl	Stored energy[a] in cavity	Stored energy[a] in half-space
0.349	3	0.886	12.4	109.3°	0.0542	0.0083
0.349	5.42	0.434	9.8[b]	100.6°	0.0422	0.0083
0.349	15	0.211	12.0	95.5°	0.0535	0.0083
0.271	3.88	2.00	71.9	134.1°	0.296	0.014
0.125	25	0.935	336	126.2°	0.682	0.057
0.125	58.3	0.371	217[b]	108.7°	0.420	0.057
0.125	80	0.293	222	105.4°	0.432	0.057

[a] The numbers are approximately proportional to stored energy.
[b] Minimum Q for this (a/λ).

[10] E. A. Guillemin, *Communication Networks* (John Wiley and Sons, Inc., New York, 1935), Vol. II, Chap. 5.
[11] S. A. Schelkunoff, Proc. Inst. Radio Engrs. **32**, 83 (1944).
[12] Montgomery, Dicke, and Purcell, *Principles of Microwave Circuits* (McGraw-Hill Book Company, Inc., New York, 1948), M.I.T. Radiation Laboratory Series, Vol. 8, Chap. 7.
[13] C. G. Montgomery, *Technique of Microwave Measurements* (McGraw-Hill Book Company, Inc., New York, 1947), M.I.T. Radiation Laboratory Series, Vol. 11, Chap. 5.
[14] J. C. Slater, Revs. Modern Phys. **18**, 467 (1946).
[15] The Q used here is the so-called "unloaded Q" of the above references.

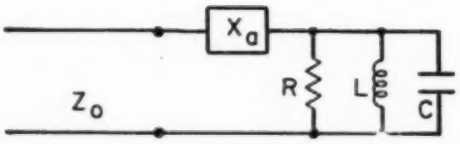


FIG. 7. Equivalent circuit near resonance.

frequency and Q. The antenna Q is related to the impedance band width on the input transmission line in exactly the same fashion as the Q of the RLC circuit is related to the band width on the transmission line in Fig. 7. In this statement the assumption is made that the feed, i.e., the method of exciting the cavity, does not appreciably perturb the cavity fields from the normal mode fields. This point is discussed below.

When the effect of the series reactance (Fig. 7) is negligible it may be shown that Q is simply related to the standing wave ratio on the transmission line. Let

$$\sigma_0=\frac{R}{Z_0}=\begin{cases}\text{SWR at resonance, if } R/Z_0>1\\ 1/\text{SWR at resonance, if } R/Z_0<1.\end{cases} \tag{20}$$

Then

$$Q=\omega_n/(\omega_2-\omega_1), \tag{21}$$

where ω_n is the resonance frequency, and ω_2 and ω_1 are the frequencies[16] at which the SWR has risen to the value r_1, where

$$r_1=\frac{(1+\frac{1}{2}\sigma_0^2+\sigma_0)^{\frac{1}{2}}+(1+\frac{1}{2}\sigma_0^2-\sigma_0)^{\frac{1}{2}}}{(1+\frac{1}{2}\sigma_0^2+\sigma_0)^{\frac{1}{2}}-(1+\frac{1}{2}\sigma_0^2-\sigma_0)^{\frac{1}{2}}}. \tag{22}$$

Experimentally, the ambiguity in σ_0 may be resolved by examining the impedance curve to determine whether R/Z_0 is greater or less than one at resonance.[12–14]

A slightly different value is obtained by including the series reactance.[14] Under the assumptions that it is constant over a wide frequency range, and that the SWR far from resonance is what would be obtained at the resonance frequency, if the resonant circuit were removed, it is possible to show that

$$r_2=\frac{[(1+\sigma_0)^2+(1+\sigma_1)^2]^{\frac{1}{2}}+[(1+\sigma_0)^2-(1+\sigma_1)^2]^{\frac{1}{2}}}{[(1+\sigma_0)^2+(1+\sigma_1)^2]^{\frac{1}{2}}-[(1+\sigma_0)^2-(1+\sigma_1)^2]^{\frac{1}{2}}}, \tag{23}$$

where r_2 is the standing wave ratio at ω_1 and ω_2. The symbol σ_0 is defined above in terms of SWR at resonance, and σ_1 is 1/SWR far from resonance. Equations (22) and (23) give similar results when $Q\gg 1$.

Experimental Results

A series of measurements were made on three square wave-guide cavity antennas. They had $l/a=7.67$, 2.00, and 0.40; with $\epsilon_r=3.80$, 3.88, and 8.41, respectively. The range of measurement covered 6 modes of the first antenna, 4 of the second, and 1 of the third. The resonance frequencies fell in the range $0.26<(a/\lambda)<0.32$.

The resonance frequencies and Q's were obtained from curves of SWR *vs* frequency, as described above. A typical curve is shown in Fig. 8. The TE_{11} mode propagates above $a/\lambda=0.358$, so apparently the asymmetry of the third and fourth modes is due to an interaction between the TE_{10} and the TE_{11} wave-guide modes. (This interaction was prohibited in the third cavity by the use of an image plane which excluded the TE_{11} mode.)

Every resonance which was due to the TE_{10} wave-guide mode alone was carefully measured, and the results are given in Tables II, III, and IV. Two values of measured Q are shown in the tables. The value obtained from Eq. (22) is called Q_1, and that from Eq. (23) is called Q_2. There are also two values of calculated Q in the tables. Q_R is the radiation Q, assuming only radiation losses, while Q_T includes dielectric losses. The first

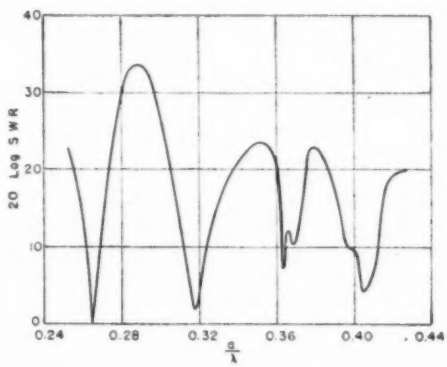


FIG. 8. SWR for $l/a=2$, $\epsilon_r=3.88$, antenna, with long probe.

two cavities were filled with sulfur for which the loss factor[17] is 0.0025. This factor is unknown for the medium in the third cavity, a lead-chloride mineral-oil mixture. The third cavity has a low Q_R and therefore dielectric losses have relatively little effect.

Effect of Feed

The theory outlined above assumes that the wave guide is uniform; but as a practical matter the cavity has to be excited, and this is accomplished through a hole in the wall, with possibly some metal protrusions. In general this feed will distort the field and change the resonance frequency. The direction of frequency shift can be found from the general perturbation theory of resonant cavities.[14] According to this theory a metal protrusion in a region of high E and low H should lower the resonance frequency, whereas one in a region

[16] This band width $(\omega_2-\omega_1)$ can be found from a Smith Chart impedance curve of the antenna. The curve is simply rotated until it coincides (as closely as possible) with a constant resistance circle. The band-width is that portion of the curve which can then be renormalized into the 2.42 SWR circle. The value of 2.42 is the minimum value of r_1.

[17] A. Von Hippel, *Tables of Dielectric Materials*, Laboratory of Insulation Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, June 1948.

of low E and high H should raise it. This effect can be seen in the tables.

The first antenna was excited by a loop in the rear wall. The loop was much smaller than the size required to produce a match at resonance and hence caused a small perturbation. The calculated and measured resonance frequencies agree almost exactly. The second and third antennas were excited by probes close to the aperture, in a region of high electric field. It was found that a match could easily be obtained by varying the probe length, and in Tables III and IV the long probe is the one which gives a match. The long probe caused a considerable perturbation which lowered the resonance frequency by as much as 8 percent. The resonance frequencies obtained with short probes were closer to the computed values, as expected.

Discussion of Q

The validity of the various assumptions which enter into the calculation are of more consequence to Q than to resonance frequency, essentially because a non-

TABLE II. Results for $l/a=7.67$, $\epsilon_r=3.80$, loop-fed, cavity antenna.

Mode	Measured a/λ	Calculated a/λ	Measured Q_1	Measured Q_2	Calculated Q_R	Calculated Q_T
1[a]	...	0.258	...	...	1470	344
2	0.263	0.264	104	134	430	220
3	0.273	0.273	89.6	114	240	156
4	0.285	0.285	82.5	97.7	172	124
5	0.301	0.300	76.9	90.4	137	105
6	0.319	0.318	67.1	75.6	119	93.4

[a] This mode was not observed.

infinite Q is a second-order effect. For example, in many cases the dielectric loss has to be included to obtain a meaningful Q, whereas its effect on resonance frequency is negligible. In some cases the metal losses also might be significant, but they were ignored in the present calculation. The measured Q values are generally less than the calculated ones, and metal losses are at least partly responsible for this.

The assumptions involved in the use of an equivalent circuit to find Q should be good when Q is high and the mode is isolated, i.e., far from its neighbors. The Q_2 values are higher than the Q_1 values and thus are closer to Q_T (except for the one case in Table III), presumably

TABLE III. Results for first two normal modes of $l/a=2.00$, $\epsilon_r=3.88$, prove-fed cavity antenna.

Mode	Probe	Measured a/λ	Calculated a/λ	Meas, ured Q_1	Measured Q_2	Calculated Q_R	Calculated Q_T
1	long	0.265	0.271	39.6	41.0	71.9	61.9
1	short	0.269	0.271	35.2	41.8	71.9	61.9
2	long	0.317	0.326	30.7	32.4	32.7	30.4
2	short	0.324	0.326	22.3	26.0	32.7	30.4

TABLE IV. Results for first normal mode of $l/a=0.40$, $\epsilon_r=8.41$, probe-fed, cavity antenna, with image plane.

Probe	Measured a/λ	Calculated a/λ	Measured Q_1	Measured Q_2	Calculated Q_R
Long	0.280	0.305	13.7	15.0	15.8
Short	0.289	0.305	11.8	14.4	15.8

because they are derived from a more elegant equivalent circuit.

The agreement obtained between measured and calculated Q is reasonable under these circumstances.

CONCLUSIONS

The normal modes of cavity antennas have been discussed and it has been shown how calculations may easily be made for them when the antenna is essentially a wave-guide cavity, and the aperture admittance is known. The experiments indicate that the theory represents the major resonance effects.

It has been demonstrated that a square wave-guide cavity antenna has a minimum Q, which is very high when a/λ is small. In view of the well-known fact that small-aperture antennas invariably have a high Q, it is felt that the minimum Q curves give the order of magnitude of the best that can be done with other shapes of cavities and other loading schemes. The minimum Q of a cavity antenna is thus determined primarily by the major aperture dimensions.

ACKNOWLEDGMENTS

This work was sponsored by the Air Research and Development Command. The author is indebted to Professor V. H. Rumsey for the stimulation of this work and for his many helpful suggestions. Credit is also due to many other members of the Antenna Laboratory for various clarifying discussions.

Multiple Shock Reflection in Corners

JOSEPH B. KELLER*
Institute of Mathematical Sciences, New York University, New York
(Received August 20, 1953)

When a shock is incident upon a corner from either the concave or convex side it is reflected by the walls forming the corner and diffracted at the edge. The mathematical problem corresponding to this phenomenon has not been solved analytically except in the following limiting two-dimensional cases: a weak (acoustic) shock incident on any corner (Keller and Blank) and a finite shock incident either normal or parallel to a nearly flat wall (Bargmann, Lighthill, Ting, and Ludloff). In this article it is pointed out that there are certain special cases of a finite shock entering a corner from the concave side in which no diffraction occurs. In these cases it is therefore possible to solve the entire problem by using the theory of regular reflection, which yields an explicit solution by algebraic means alone.

I. INTRODUCTION

WHEN a shock enters a corner it is reflected one or more times from the walls forming the corner. Finally it reaches the corner and then, generally, it is diffracted. It is the purpose of the present article to point out that in special cases no such diffraction occurs. In these cases it is therefore possible to solve the entire reflection problem exactly and explicitly by algebraic means only.

To exemplify this result the solutions are found for several particular cases. These solutions show that the pressure on the walls increases tremendously because of the multiple reflections. It therefore seems that such corners are very susceptible to damage from explosion shocks, and should be avoided in designing structures likely to be exposed to such shocks.

The fact that solutions without diffraction can occur is indicated by the results for acoustic pulses entering corners. These results are given in "Diffraction and Reflection of Pulses by Wedges and Corners" by A. Blank and the present author [Communs. Pure and Appl. Math. **IV**, No. 1 (June, 1951)]. In that article it is shown that in the two-dimensional case with plane pulses incident there are certain corner angles $\pi/2n$ (n any integer) for which no diffraction occurs. This result holds for any angle of incidence of the acoustic shock, and these angles are just those for which every optical ray is reflected back parallel to itself.

It is therefore natural to seek corresponding cases in which finite shocks are reflected without diffraction. Just as in the acoustic problem, such cases arise when that portion of the shock front which suffers its final reflection from one wall emerges parallel to that portion which suffers its final reflection from the other wall. If such cases occur, and if all reflections are regular, the flow field will be piecewise constant and obtainable by algebraic calculations, i.e., by calculating a finite number of regular reflections.

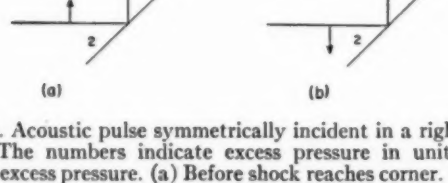


FIG. 1. Acoustic pulse symmetrically incident in a right-angled corner. The numbers indicate excess pressure in units of the incident excess pressure. (a) Before shock reaches corner. (b) After shock reaches corner.

II. THE RIGHT-ANGLED CORNER

In order to construct an example of such a case, let us begin by examining the acoustic solution for a plane shock incident symmetrically in a right-angled corner with rigid walls. The solution at two typical instants is shown in Fig. 1, the numbers indicating excess pressure in units of the excess pressure of the incident shock. This solution is given by Courant and Friedrichs,[1] and it is almost obvious.

Let us seek a finite shock solution in which the shock configuration is the same as that of Fig. 1, but in which the pressure values are necessarily different. If we indicate the state of the gas in each region in Fig. 1 by one of the numbers 0, 1, 2, 3, 4, then we may designate the corresponding pressures by p_0, p_1, p_2, p_3, p_4. In Fig. 1 the angle of incidence of each shock with respect to the wall from which it is reflected is $\pi/4$ ($=45°$) and each angle of reflection is also $\pi/4$ ($=45°$). Now in general the angles of incidence and reflection are unequal in finite shock reflection. However for one particular angle α in each gas they are equal. In a polytropic gas with adiabatic exponent γ this angle α is given by (reference 1, p. 327)

$$\cos 2\alpha = \frac{\gamma-1}{2}. \tag{1}$$

If $\alpha=45°$, as in Fig. 1, then $\gamma=1$. Thus the configuration of Fig. 1 is possible with an incident finite shock of any strength provided $\gamma=1$.

To find the pressures, we assume that p_0 and p_1 are given. Then p_2 is determined by the conditions at a

* On leave of absence during 1953–4 at the Office of Naval Research, Washington, D. C.

[1] R. Courant and K. O. Freidrichs, *Supersonic Flow and Shock Waves* (Interscience Publishers, Inc., New York, 1948).

regular reflection which yield when $\gamma=1$ (reference 1, pp. 331, 327)

$$p_2=\frac{p_1}{p_0}p_1. \tag{2}$$

To find p_3 we note that it is the pressure resulting from the head-on collision of two shocks and this is equivalent to a regular normal reflection. Thus

$$p_3=\frac{(p_2)^2}{p_1}=\left(\frac{p_1}{p_0}\right)^2 p_1. \tag{3}$$

Finally, p_4 results from another regular reflection so

$$p_4=\frac{(p_3)^2}{p_2}=\left(\frac{p_1}{p_0}\right)^3 p_1. \tag{4}$$

All other quantities can be found from the formulas in reference 1, p. 330. The preceding results also apply to the corner of angle $\pi/4$ obtained by bisecting Fig. 1.

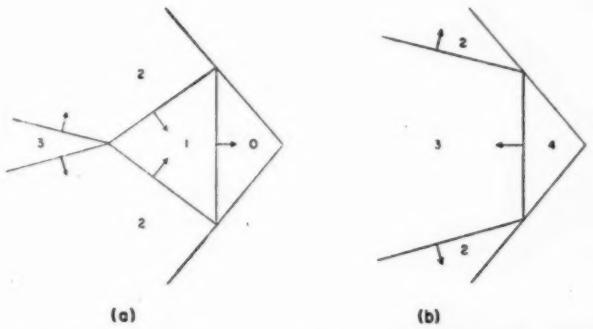


FIG. 2. Shock configurations which may result from a finite shock symmetrically incident in a corner. Cases in which this occurs are shown in Fig. 4. The numbers denote the regions. (a) Before shock reaches corner. (b) After shock reaches corner.

It is interesting to note that p_4 is greater than p_1 by the factor $(p_1/p_0)^3$ which equals 1000 if $p_1/p_0=10$. Thus p_4 may become very large. This fact must be approximately true for other values of γ and other corner angles as well.

For other values of γ the configuration in Fig. 1 is not possible, even with any other corner angle. This is so because in Fig. 1 the angles of incidence and reflection are complementary for any corner angle. Furthermore the angle of reflection in Fig. 1(a) is the angle of incidence in Fig. 1(b). But when the angles of incidence and reflection are complementary, the latter must always exceed the former (see, e.g., reference 1, p. 328, Fig. 55). This cannot apply in both cases unless the angles of incidence and reflection are equal, which was the case treated above.

Thus to obtain solutions for $\gamma\neq1$ we must consider configurations such as those in Figs. 2 and 3 rather than Fig. 1. To analyze either of these cases we must specify the corner angle, the state 0, the state 1, and the direction of incidence, which is along the angle bisector

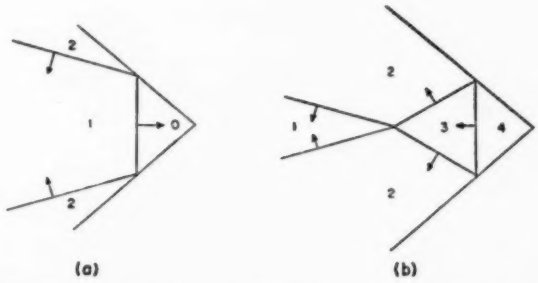


FIG. 3. Shock configurations which may result from a finite shock symmetrically incident in a corner. No cases in which this occurs have been found in this paper. (a) Before shock reaches corner. (b) After shock reaches corner.

in the figures. Then we must determine state 2 by regular reflection, state 3 by regular reflection from state 2, and state 4 by regular reflection from state 3. This analysis will yield a solution provided that the last shock between states 3 and 4 is perpendicular to the bisector of the corner angle (for symmetric incidence), since only then will the two halves of this shock form a straight line. Thus the specified data must be adjusted until this last condition is satisfied.

If the gas, the state 0, and the direction of incidence are fixed, then the incident shock strength and the corner angle remain adjustable. It seems that these two variables can always be adjusted so as to yield an explicit solution of the type considered above. In fact for either of the variables fixed in an appropriate range, it seems that a solution can be found by adjusting the other. Therefore in the plane in which the incident pressure ratio p_1/p_0 and the corner angle θ are coordinates, there is a curve corresponding to diffractionless reflection. As p_1/p_0 tends to unity, this curve must tend to $\theta=\pi/2$ for any γ and any direction of incidence, as the acoustic solution shows. For symmetric incidence and $\gamma=1$ this curve is just the vertical line $\theta=\pi/2$. For symmetric incidence and $\gamma\neq1$ the curve is shown in Fig. 4 starting at the acoustic point $p_1/p_0=1$, $\theta=\pi/2$. When $\gamma>1$, then $\theta>\pi/2$ as the curve shows. Of course there are undoubtedly other branches of this curve corresponding to the other angles $\pi/2n$ which occur in the acoustic case. The method by which this curve was obtained is described in the next section.

III. CALCULATION FOR $\gamma\neq1$

In order to find the curve in the $(p_0/p_1-\theta)$-plane which gives diffractionless reflection, we restrict our attention to symmetric incidence and the configuration shown in Fig. 2. We assume that the incident pressure ratio $\xi^{(1)}=p_0/p_1=1+\eta$ where η is small and negative. We also assume that $\theta=(\pi/2)-a\eta^2+\cdots$ where a is to be determined. This form of θ was decided upon after examining the equations. Now we follow reference 1, pp. 329–330, setting $t_0^{(1)}=\tan\alpha_0^{(1)}$, $t_2^{(1)}=\tan\alpha_2^{(1)}$. Three superscripts 1, 2, 3 will be used, referring to the first, second, or third reflection, respectively. Then we have (reference 1, p. 330, Eqs. 127.04, 127.08) with $(\gamma-1)/$

$(\gamma+1)=\mu^2$

$$\frac{(1-\xi)t_0}{1+\mu^2\xi=(\xi+\mu^2)(t_0)^2}=\frac{(\zeta-1)t_2}{1+\mu^2\zeta+(\zeta+\mu^2)(t_2)^2}=M \quad (5)$$

$$M^2(1-\mu^2)^2(t_0-t_2)+M\{(1-\mu^2)^2-(t_0-t_2)^2 - (\mu^2+t_0t_2)^2\}-(t_0-t_2)=0. \quad (6)$$

With the chosen values of $\xi^{(1)}$ and θ we find $\alpha_0{}^{(1)}=(\pi/2)-(\theta/2)$ from the figure, compute $M^{(1)}$ from (5), solve (6) for $t_2{}^{(1)}$, then solve (5) for $\zeta^{(1)}=p_2/p_1$. Now $\alpha_0{}^{(2)}=\alpha_2{}^{(1)}-(\theta/2)$ and $t_0{}^{(2)}=\tan\alpha_0{}^{(2)}$ so we compute $M^{(2)}$ from (5) using $\xi^{(2)}=1/\zeta^{(1)}$. Then we solve (6) for $t_2{}^{(2)}$ and (5) for $\zeta^{(2)}$. Again from the figure $\alpha_0{}^{(3)}=(\theta/2)-\alpha_2{}^{(2)}$, $t_0{}^{(3)}=\tan\alpha_0{}^{(3)}$. Then we compute $M^{(3)}$ from (5) with $\xi^{(3)}=1/\zeta^{(2)}$, solve (6) for $t_2{}^{(3)}$ and (5) for $\zeta^{(3)}$. Finally we require that $\alpha_2{}^{(3)}=\alpha_0{}^{(1)}$ or $t_2{}^{(3)}=t_0{}^{(1)}$. This condition determines θ in terms of $\xi^{(1)}$, or in the perturbation problems it yields the constant a. The results of this calculation are, to second order in η,

First Reflection

$$\xi^{(1)}=1+\eta \quad (7)$$

$$\theta=\frac{\pi}{2}-a\eta^2 \quad (8)$$

$$t_0{}^{(1)}=1+a\eta^2 \quad (9)$$

$$M^{(1)}=\frac{-\eta}{2(1+\mu^2)}+\frac{\eta^2}{4(1+\mu^2)} \quad (10)$$

$$t_2{}^{(1)}=1-\frac{2\mu^2\eta}{(1+\mu^2)}+\left[a+\frac{3\mu^2}{1+\mu^2}\right]\eta^2 \quad (11)$$

$$\zeta^{(1)}=1-\eta+\eta^2. \quad (12)$$

Second Reflection

$$t_0{}^{(2)}=-\frac{\mu^2\eta}{1+\mu^2}+\left[a+\frac{\mu^2(3+\mu^2)}{2(1+\mu^2)^2}\right]\eta^2 \quad (13)$$

$$\xi^{(2)}=1+\eta+0\cdot\eta^2 \quad (14)$$

$$M^{(2)}=\frac{\mu^2\eta^2}{(1+\mu^2)^2} \quad (15)$$

$$t_2{}^{(2)}=-\frac{\mu^2\eta}{1+\mu^2}+\left[a+\frac{\mu^2(1+5\mu^2)}{2(1+\mu^2)^2}\right]\eta^2 \quad (16)$$

$$\zeta^{(2)}=1-\eta-\left[\frac{1+\mu^2}{\mu^2}a+\frac{1+3\mu^2}{2(1+\mu^2)}\right]\eta^2. \quad (17)$$

Third Reflection

$$t_0{}^{(3)}=1+\frac{2\mu^2}{1+\mu^2}\eta-\left[3a+\frac{\mu^2(1+3\mu^2)}{(1+\mu^2)^2}\right]\eta^2 \quad (18)$$

$$\xi^{(3)}=1+\eta+\left[\frac{1+\mu^2}{\mu^2}a+\frac{5\mu^2+3}{2(\mu^2+1)}\right]\eta^2 \quad (19)$$

$$M^{(3)}=-\frac{\eta}{2(1+\mu^2)}-\left[\frac{a}{\mu^2}+\frac{2\mu^2+1}{2(1+\mu^2)^2}\right]\eta^2 \quad (20)$$

$$t_0{}^{(3)}=1+0\cdot\eta-\left[5a+\frac{9\mu^4+5\mu^2}{(1+\mu^2)^2}\right]\eta^2 \quad (21)$$

$$\zeta^{(3)}=1-\eta-\left[\frac{1+\mu^2}{\mu^2}a+\frac{3\mu^2+1}{2(1+\mu^2)}\right]\eta^2. \quad (22)$$

In order that $t_2{}^{(3)}=t_0{}^{(1)}$ we find that a in Eq. (8) must be given by

$$a=-\frac{9\mu^4+5\mu^2}{6(1+\mu^2)^2}=\frac{(1-\gamma)(7\gamma-2)}{12\gamma^2}. \quad (23)$$

We see that $a=0$ at $\gamma=1, 2/7$, is positive between these values and negative outside them. Since usually $\gamma>1$, we see that θ must be increased above $\pi/2$ as the incident shock strength is increased, in order to yield reflection without diffraction. We note that if $\gamma=1$, $a=0$ and θ need not change, in agreement with the result of the previous section for $\gamma=1$. For $\gamma=1.4$ we find $a=-13/296$. Some of these results are shown in Fig. 4.

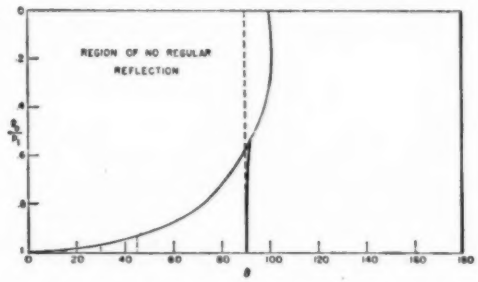


FIG. 4. Curves showing conditions in which reflection without diffraction occurs as shown in Figs. 1 and 2. The pressure ratio across the incident shock is p_0/p_1, the corner angle is θ, and the shock is incident symmetrically. The acoustic points $p_0/p_1=1$, $\theta=\pi/2n$ with n any integer were found previously (see Sec. I). The line $\theta=\pi$ corresponds to normal reflection. The dashed line $\theta=\pi/2$ (- - -) is one branch of the curve for $\gamma=1$ and was determined in Sec. II. The solid curve starting at $p_0/p_1=1$, $\theta=\pi/2$ was computed from Eq. (8) with the constant $a=-13/296$ calculated from Eq. (23) using $\gamma=1.4$ (air). The curve separating the region of regular reflection from that in which no regular reflection occurs is given by the equation $\theta_{ext}=\pi-2\alpha_{ext}$ where α_{ext} is taken from [1, p. 326, Fig. 54] and is computed for $\gamma=1.4$ (air). The dotted branches of the curves are sketched in, but have not been computed.

Finally we note that the computational procedure outlined above can also be used for shocks that are not nearly acoustic. In that case an incident shock strength must be chosen as well as a definite corner angle. If these choices do not yield a solution one of them (e.g., the angle) must be changed and the calculation repeated until a solution is obtained. However since the range of shocks for which reflection without diffraction occurs is limited to shocks of small strength, as is shown in Fig. 4, the preceding results should be adequate, and further calculation seems unnecessary.

Measurement of the Effect of Chlorine Treatment on the Work Function of Titanium and Zirconium*

HERBERT MALAMUD AND AARON D. KRUMBEIN
Physics Department, University of Maryland, College Park, Maryland
(Received August 20, 1953)

This paper describes work function measurements of titanium and zirconium carried out by the conventional photoelectric method. Two samples of each metal were investigated before and after treatment with chlorine. The changes in work function due to chlorine absorption were of the same order of magnitude as those reported for other halogens on other metals.

INTRODUCTION

WORK function measurements were made by the photoelectric method[1] on two samples each of titanium and zirconium before and after treatment with chlorine. It is well known that most gases occluded on a metal surface generally lower the work function (ϕ) while a few gases, notably the halogens, raise the work function. The present work was done as part of a research project on halogen counters to determine whether the metals under test would be suitable as counter cathode materials.[2] The metal samples and their analysis were supplied by the United States Bureau of Mines at College Park, Maryland. The titanium was commercially pure, arc-melted metal with 0.04–0.06 percent carbon, less than 0.03 percent nitrogen, and less than 0.1 percent oxygen. The zirconium used was "hafnium-containing, induction melted," with somewhat more than 0.2 percent carbon.

APPARATUS

The experimental arrangement was straightforward. An ultraviolet lamp (G.E. Type CH-3) illuminated the input slit of a small Gaertner quartz prism spectrometer which had been fitted with a movable slit (in place of the film plate), and a scale, calibrated for the identification of the various ultraviolet spectral lines.

Light from each spectral line in turn selected by this monochromator was allowed to enter the quartz window of a Pyrex testing tube, and illuminate the cylindrical cathode being tested. The anode was a tungsten wire coaxial with the cathode.

The tube was supported in a brass case with the quartz end window projecting through a hole punched in a sheet of rubber forming one end of the case. The inside of the case contained the first stage of a dc amplifier and suitable drying agents. The circuit of the amplifier was identical with the Victoreen 247B radiation survey meter circuit except that our testing tube was substituted for the ionization chamber. A Leeds and Northrup Type R galvanometer served as the indicator. Figure 1 shows a block diagram of the apparatus.

PROCEDURE AND RESULTS

Each sample was tested exactly as received and again after degassing. The first degassing process was accomplished by heating the sample to bright red heat (~900°C) in vacuum by means of a radio frequency heater for about ten minutes. The vacuum system consisted of a two-stage water-cooled DPI oil diffusion pump, with a dry ice-alcohol trap between the pump and the sample. A pressure of less than 10^{-4} cm Hg (the limit of the thermocouple gauge used), was obtained even while the degassing was in progress. The metal samples were then further degassed and after cooling, were exposed to dry chlorine gas at a pressure of 10 mm Hg for the first sample of each metal, and 20 mm Hg for the second sample. "Lecture bottle" chlorine gas, dried over phosphorus pentoxide, was used. The metals were then allowed to "pickle" in the chlorine for about eighteen hours. Measurements were made after reevacuation at the end of that time.

The final treatment, which we call "heat pickling," was accomplished by admitting the chlorine while the metal was at red heat immediately after degassing. The metal was then allowed to cool in the presence of the chlorine. The first samples were so treated with chlorine at a pressure of about 4 mm of mercury and almost all of it was absorbed by the metal. The second samples were treated with chlorine at a pressure of

TABLE I. Work function of titanium and zirconium metals before and after treatment with chlorine.

Samples No. 1	Zirconium	Titanium
Sample as received	3.10±0.07 ev	3.10±0.07 ev
Degassed	4.33±0.07	4.45±0.05
Pickled	4.76±0.11	4.76±0.11
Heat pickled	4.72±0.07	5.54±0.06
Samples No. 2	**Zirconium**	**Titanium**
Sample as received	3.81±0.12 ev	4.36±0.04 ev
Degassed	4.33±0.07	4.45±0.05
Pickled	4.76±0.11	4.76±0.11
Heat pickled	4.72±0.07	5.27±0.09

* This work is supported by the Bureau of Ships under Contract NOBsr-49066.

[1] A. L. Hughes and L. A. Dubridge, *Photoelectric Phenomena* (McGraw-Hill Book Company, Inc., New York, 1932), pp. 419-456.

[2] D. H. Wilkinson, *Ionization Chambers and Counters* (Cambridge University Press, Cambridge, 1950), pp. 44–45.

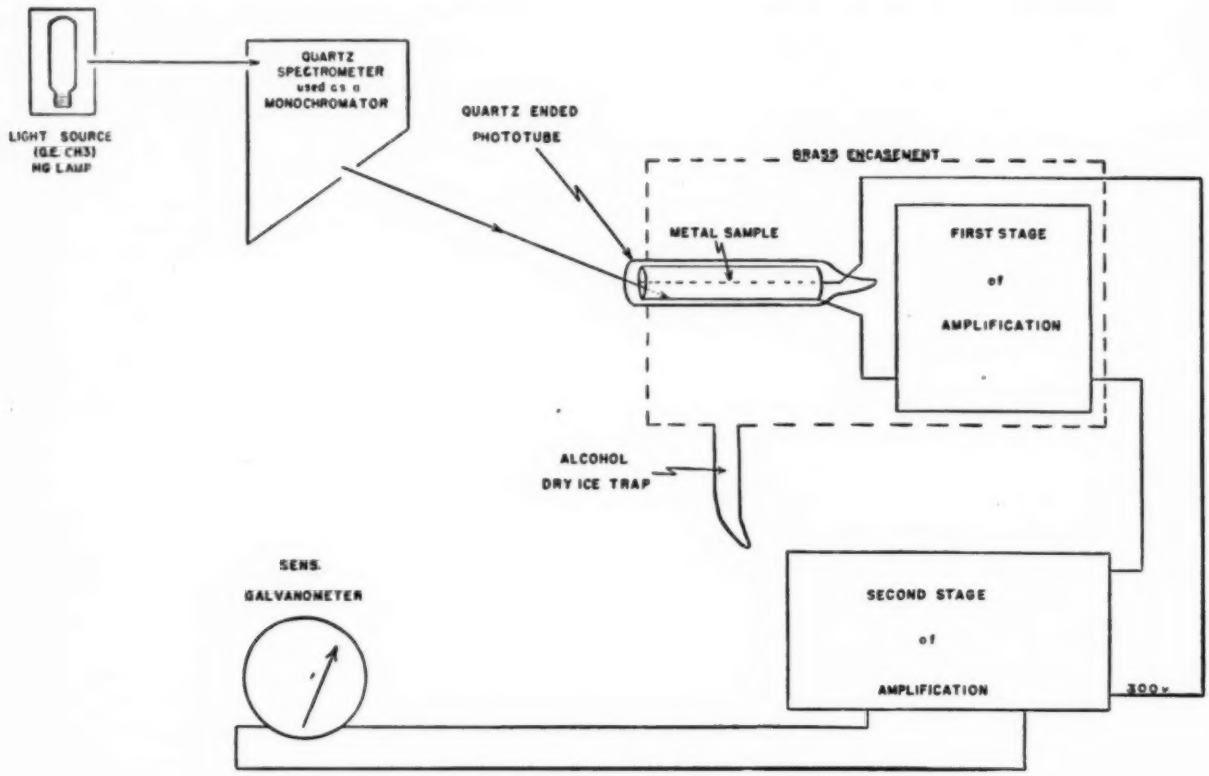


FIG. 1. Block diagram of the apparatus.

about 20 mm of mercury of which about half was absorbed. The second titanium sample was not preheated to as high a temperature and subsequently did not absorb the chlorine to the same extent as did the other sample. Its work function, as will be seen in Table I, was substantially lower than that of the first titanium sample.

Each of these measurements was performed at least twice before the results were accepted. Although a period of several months often elapsed between tests of samples 1 and 2, reproducible results were obtained as can be seen in Table I.

DISCUSSION OF RESULTS

The limits of accuracy of the measurements were set by the spacing of the mercury spectral lines. Also, we believe that each process of treatment was carried to its limit. For example, in each degassing process, the sample was heated until some of the cathode metal was evaporated onto the glass containers. Further, it was noted that although different amounts of chlorine were used for each sample in the pickling processes, the values of ϕ are the same, indicating that most probably the same amount of chlorine was adsorbed in each case.

Ouellet and Rideal,[3] measuring work function by using the metal in question as the cathode of a Geiger tube, found that the shift in ϕ due to a gas is proportional to the amount of gas adsorbed. They measured the effect of up to 0.2 mm Hg of iodine and bromine vapor on the work function of gold, silver, and tungsten Geiger tube cathodes. The maximum changes in work function, found by them to be due to halogen gases range from 0.21 ev to 0.40 ev. Our own results for the pickling process, when more than sufficient chlorine was present for saturation, according to the standards of Ouellet and Rideal, gave us, for admittedly different metals and a different halogen, the value range 0.47 ev $\leq\Delta\phi\leq$0.61 ev. These are the maximum values.

The work functions of the degassed metals may be compared with the values obtained by Rentschler and Henry.[4] They found, for titanium, a value of ϕ of 4.17 ev and for zirconium, a value of ϕ of 4.0 ev. They made their tests on pure titanium and pure zirconium deposited by sputtering on thin wires. This may explain the fact that their values of the work function were slightly lower than our own.

[3] G. Ouellet and E. K. Rideal, J. Chem. Phys. **3**, 150 (1935).
[4] H. C. Rentschler and D. E. Henry, Trans. Electrochem. Soc. **87**, 289 (1945).

Delayed Yield and Strain Rate and Temperature Dependence of Yield Point in Iron

TAKEO YOKOBORI
Institute of Science and Technology, University of Tokyo, Komaba-machi, Meguro, Tokyo, Japan
(Received September 1, 1953)

A unified explanation of delayed yield and of strain rate and temperature dependence of yield point in iron was attempted on the basis of the Cottrell locking theory, taking internal stress into account. The energy of a dislocation required to fit existing data is small by an order of magnitude compared with the value theoretically estimated by Mott and Nabarro. The distance, calculated from the value of internal stress, between dislocations assembled at a grain boundary is of a reasonable order of magnitude. It is suggested that the mechanism of yielding may change from Cottrell type to Mott-Nabarro or Frank-Read type when applied stress falls below some limiting value.

I. DELAYED YIELD

WOOD and Clark[1] carried out experiments upon delayed yield, that is, the phenomena of the time delay of yielding under constant stress in mild steel. In this paper it is attempted to explain their measurements in terms of the present theory of dislocations. On the basis of the dislocation locking theory developed by Cottrell,[2] the mean delay time τ for yielding is a reasonably good approximation[3] given by the form

$$\tau = 1/m = \tau_0(\sigma/\sigma_0)^{-(1/nkT)}, \tag{1}$$

where

m = the rate of breaking away of dislocation, i.e., $md\tau$ is the probability per time interval, $d\tau$ that a dislocation breaks away.
$1/\tau_0 = Z\nu$.
ν = the vibration frequency of a dislocation.
Z = the number of dislocations for breaking away in the region of stress concentration such as shoulders of a specimen.
σ = applied constant tensile stress.
σ_0 = tensile yield stress at absolute zero temperature.
$1/n$ = the constant[4] with the dimension of energy which is a function of A, W, ρ and λ. The definitions of A, W, ρ and λ are the same as in the paper of Cottrell and Bilby[5] (C–B).
k = the Boltzmann's constant.
T = absolute temperature.

If we take account of internal stress,[6] σ_i,[7] arising from the presence of other nearby dislocations assembled at a grain boundary as an array of dislocations, Eq. (1) can be written as

$$\tau = \tau_0(\sigma - \sigma_i/\sigma_0 - \sigma_i)^{-(1/nkT)}, \tag{2}$$

where $\sigma - \sigma_i = \sigma_e$ = applied constant tensile stress for breaking away dislocations anchored by Cottrell locking, and $\sigma_0 - \sigma_i = \sigma_{0e}$ = break-away stress by Cottrell at absolute zero temperature. Assuming $A = 5\times10^{-21}$ dyne·cm², $W = 0.1$ ev per atom plane, $\rho = 2\times10^{-8}$ cm, $1/\tau_0 = 1.3\times10^{9}$ sec^{-1}, $\sigma_{0e} = 158$ kg/mm², and $\sigma_i = 18.8$ kg/mm², the theoretical dependence of delay time upon stress and temperature is in good agreement[8] with the experiments of Wood and Clark[1] as shown in Fig. 1.

II. STRAIN RATE AND TEMPERATURE DEPENDENCE OF YIELD POINT

Yield stress under constant strain rate $\dot{\epsilon}$ is also given as a good approximation,[3] on the basis of the same consideration given to delayed yield,

$$\sigma - \sigma_i = (\sigma_0 - \sigma_i)(\dot{\epsilon}E\tau_0/nkT\sigma_0)^{nkT}, \tag{3}$$

where σ = the observed upper (or sharply flowing) tensile yield stress and E is Young's modulus. The data referred to below are assumed to satisfy this condition of constant strain rate. It can be seen from curve I in Fig. 2 that the theory is roughly in agreement with the experiments of Wood and Clark[1] when the same values of A, W, ρ, τ_0, σ_{0c}, and σ_i as in Sec. I are used and $\dot{\epsilon}$ is

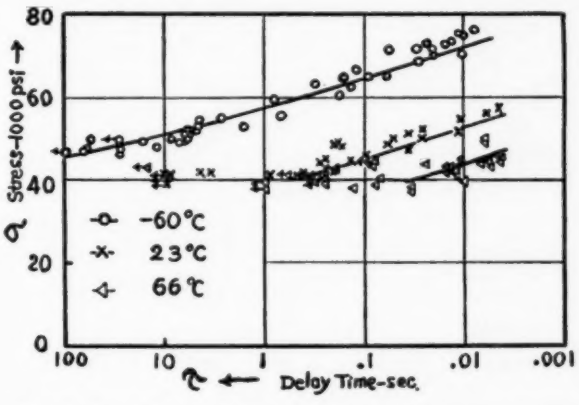


FIG. 1. Dependence of delay time for yielding upon stress and temperature. Solid curves are calculated by Eq. (2). From data of Wood and Clark (see reference 1).

[1] D. S. Wood and D. S. Clark, Trans. Am. Soc. Metals 43, 571 (1951).
[2] A. H. Cottrell, Report of a Conference on the Strength of Solids, Physical Society (London), p. 30 (1948).
[3] T. Yokobori, Phys. Rev. 88, 1423 (1952).
[4] $1/n$ is not exactly expressed by $\alpha(AW\rho)^{\frac{1}{2}}$ in the case where W or σ/σ_0 is very small.
[5] A. H. Cottrell and B. A. Bilby, Proc. Phys. Soc. (London) 62, 19 (1949).
[6] A. H. Cottrell, Symposium on the Plastic Deformation of Crystalline Solids, U. S. Office of Naval Research, p. 60 (1950).
[7] σ_i is the mean value of the internal stress corresponding to the position x_1 of stable equilibrium and x_2 of the unstable one in the C–B paper (see reference 5).
[8] The data at 121°C are too few to be compared with the theory.

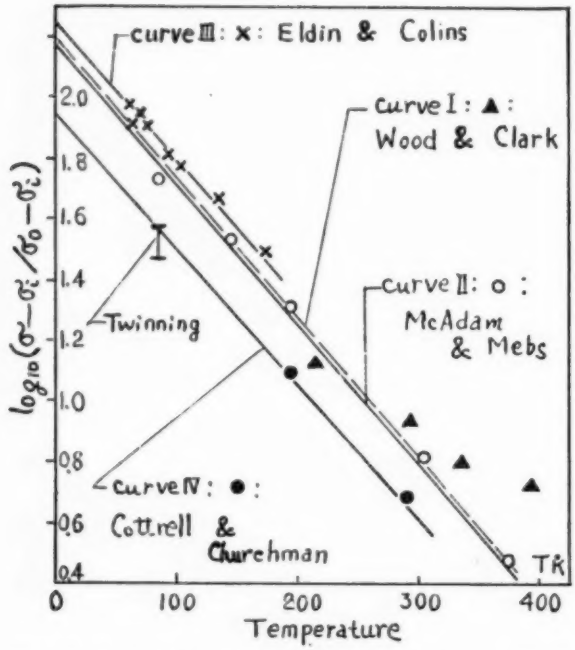


FIG. 2. Dependence of yield point upon temperature. Solid curves are calculated by Eq. (3). Constants used are shown in Table I.

assumed to be 3.6×10^{-5} sec^{-1}. It can also be understood from curves II, III, and IV in Fig. 2 that other data[9-11] available from literature with iron and mild steel are in good accordance with the theory when different values of σ_{0c}, and σ_i from, but same values of A, W, ρ_0, τ_0, and $\dot{\epsilon}$ as those in curve I are assumed (Table I).

III. DISCUSSION

In this paper a single break-away process was assumed to be the rate-determining event in the initiation of yielding.

It is noted that the energy of a dislocation required to fit existing data is small by an order of magnitude compared with the value estimated theoretically by Mott and Nabarro.[12]

TABLE I. Values of constants in Eqs. (2) and (3), stress concentration factor q, and dislocation distance "a."

($A=5\times10^{-21}$ dyne cm^2, $W=0.1$ ev per atom plane, $\rho=2\times10^{8}$ cm, $1/\tau_0=1.3\times10^{9}$ sec^{-1}, $\dot{\epsilon}=3.6\times10^{-5}$ sec^{-1}.)

Investigators	$\sigma_{0c}=\sigma_0-\sigma_i$ kg/mm^2	σ_i kg/mm^2	q	a 10^{-8} cm
Wood and Clark	158	18.8	1.7	0.39
McAdam and Mebs	151	8.2	1.8	0.58
Eldin and Collins	178	11.0	1.5	0.55
Cottrell and Churchman (single crystal)	90	0	3.0	

If the region of stress concentration involves only the surface area of shoulders of a specimen, say of the order of 1 cm^2, then Z is assumed to be about 10^8 in annealed state. From the value of $1/\tau_0$ fitted to existing data, we find $\nu\simeq10$ sec^{-1}, which is very much smaller than that estimated by Mott and Nabarro.[12]

The theoretical break-away stress, σ_{t0}, for anchored dislocation at absolute zero temperature becomes 270 kg/mm^2, when calculated from $3\sqrt{3}A/4\lambda^2\rho^2$, which is greater than the value of $\sigma-\sigma_i(=\sigma_{0c})$ extrapolated to 0°K. This may be due to stress concentration, a factor of which is given by $q=\sigma_{t0}/\sigma_{0c}$ (Table I).

Assuming that the dislocation which is attempting to break away is a member of an infinite wall of anchored dislocations with the same sign, a distance "a" between dislocations can be calculated[6] from $q\sigma_i=G\lambda\pi^2\Delta\chi/2\pi(1-\nu)9a^2$. The value of "$a$" calculated is shown in Table I, which is slightly smaller than 10^{-6} cm but of a reasonable order of magnitude. It can be seen from this analysis that q is larger and σ_i is far smaller (almost negligible) in single crystal than in polycrystal. The negligible order of σ_i in single crystal means that grain boundary acts as an obstacle.

The calculated ratio of the activation energy to kT, i.e., U/kT is found to be about 23, which is smaller than 50 and of a reasonable order.

When applied stress falls below some limit, the mechanism of yielding is expected to change from Cottrell type to Mott-Nabarro type in the presence of internal stress, or to Frank-Read type. The discrepancy between the experiments and the theory as shown in Fig. 1 seems rather to be the result of this transition than the result of rough approximation in the range of very small $\sigma-\sigma_i/\sigma_0-\sigma_i$. It seems that because of the above transition the yield stress becomes almost independent of strain rate[13] when the latter becomes slower, that is, stress becomes smaller than some limiting value. It also seems to be related to this transition that in the experiments of Kramer and Maddin[14] with β brass, the stress for initiation of plastic flow does not appear to be temperature and time dependent in the range of small applied stress. Seitz[15] attempted to give these related phenomena another interpretation.

On the basis of the suggestion put forward by Nabarro,[16] the dependence of the delay time upon applied stress and temperature seems difficult to discuss quantitatively in his original idea.

The author wishes to express his sincere thanks to Professor D. S. Clark and Professor A. H. Cottrell for sending him reprints of their papers. This work was carried out by the financial aids of the Ministry of Education of Japan.

[9] D. J. McAdam, Jr. and R. W. Mebs, Proc. Am. Soc. Testing Materials **43**, 661 (1943).
[10] A. E. Eldin and S. C. Collins, J. Appl. Phys. **22**, 1296 (1951).
[11] A. H. Cottrell and A. T. Churchman, Nature **167**, 943 (1951).
[12] N. F. Mott and F. R. N. Nabarro, Report of a Conference on the Strength of Solids, Physical Society (London), p. 1. (1948).
[13] E. A. Davis, Trans. Am. Soc. Mech. Engrs. (J. Appl. Mech.) **A137**, 60 (1938).
[14] I. R. Kramer and R. Maddin, J. Metals **4**, 197 (1952).
[15] F. Seitz, Phys. Rev. **88**, 722 (1952).
[16] F. R. N. Nabarro, Report of a Conference on the Strength of Solids, Physical Society (London), p. 38 (1948).

JOURNAL OF APPLIED PHYSICS VOLUME 25, NUMBER 5 MAY, 1954

Negentropy and Information in Telecommunications, Writing, and Reading

L. BRILLOUIN
I. B. M. Watson Laboratory, New York, New York
(Received September 8, 1953)

The negentropy principle of information was discussed in a previous paper and applied to a variety of problems of scientific observation. The present paper considers the technical problems of telecommunication. Using a thermodynamical discussion one obtains the Shannon formula for the capacity of a channel with noise. This example proves again the connection between information and negentropy.

The problem of writing and reading is examined. Here we have a type of "dead information" which must be re-energized for reading: The source of light needed for reading provides the negentropy which is changed into information by the reader.

The negentropy connected with "organization" is also discussed. All these examples show the connection between information and negentropy. They are of great importance for a theory of information, although the entropy amounts involved may be completely negligible.

1. THE NEGENTROPY PRINCIPLE OF INFORMATION

THE statistical definition of information was discussed in a previous paper.[1] The problem of information connected with a physical system was investigated and it was found that "information," in such a case, always corresponds to a negative term in the total entropy of the system (negentropy).

Special emphasis was laid on the conditions of observation in the physical sciences. In a variety of examples it was proved that an experiment, by which an information is obtained about a physical system, corresponds in average to an increase of entropy in the system or in its surrounding. An information must always be paid for in negentropy, the decrease of negentropy being larger than (or equal to) the amount of information obtained.

A generalized Carnot's principle was stated:

"The sum of negentropy plus information must always decrease."

The present paper will discuss some applications of these general ideas to some practical questions of everyday life: telecommunication, speech, writing, and reading. It will be proved, in the first examples, that the negentropy principle of information leads directly to the well-known formula of C. E. Shannon for the capacity of a communication channel. The problem of communication through writing and reading requires a special discussion that will bring forward some new points of real importance.[2]

The problems investigated in the previous paper corresponded to a variety of examples of transformations of the general character

$$\text{negentropy} \rightarrow \text{information}. \tag{1}$$

The telecommunication channel presents a typical case of the inverse transformation

$$\text{information} \rightarrow \text{negentropy} \tag{2}$$

at the transmitting end. On the contrary, the receiving end changes negentropy back into information.

Writing and reading provide examples of both transformations (1) and (2).

2. TELECOMMUNICATION ON A CHANNEL WITH NOISE

The problem of the telecommunication channel will be discussed along the lines of a paper of Raymond[3] and some results of Felker and Bell[4] will be obtained. We start with Shannon's fundamental formula for the capacity of a channel with noise.[5] We consider, for instance, a cable which can transmit all the frequencies from 0 up to a maximum ν_M. The cable is at a temperature T and consequently exhibits a noise level corresponding to a certain average power P_n. When the band width ν_M *and the direction of propagation* of the signals are given, we obtain, at a temperature T

$$P_n = kT\nu_M. \tag{3}$$

This results directly from the classical discussion of Nyquist about the Johnson noise.[6]

At the transmitting end, a transmitter emits signals with an average power P. The information to be transmitted along the cable is translated into a system of pulses or in a variety of frequencies which travel along the cable. Shannon proves that the maximum capacity of the channel is given by the formula

$$C = K\nu_M \ln\left(1+\frac{P}{P_n}\right). \tag{4}$$

If we choose $K=\log_2 e$, the capacity is given in "bits per

[1] L. Brillouin, J. Appl. Phys. **24**, 1152 (1953).
[2] L. Brillouin, Am. Scientist **38**, 594 (1950); see also a paper in the book L. de Broglie, *Physicien et Penseur* (Albin Michel, Paris, 1953), p. 359.
[3] R. C. Raymond, Am. Scientist **38**, 289 (1950). This paper contains a wrong coefficient 4 in some formulas.
[4] J. A. Felker, Proc. Inst. Radio Engrs. **40**, 728 (1952); D. A. Bell, Am. Scientist, **40**, 682 (1952).
[5] C. E. Shannon and W. Weaver, Mathematical Theory of Communication (University of Illinois Press, Urbana, 1949), p. 67, Th. 17; Meyer-Leifer and W. F. Schreiber, *Advances in Electronics* (Academic Press, Inc., New York, 1951), vol. III, p. 306.
[6] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals* M.I.T. Radiation Laboratory Series, **24** (McGraw-Hill Book Company, Inc., New York, 1950), p. 66, first line.

second." Whatever coding system we use, the channel cannot transmit more than C bits of information per second.

Instead of K we shall use in formula (4) the Boltzmann constant k and measure the information in thermodynamical units.

3. THE THERMODYNAMICS OF A CHANNEL WITH NOISE

The cable is initially at a temperature T and obtains, under normal conditions, an entropy S_0. When we transmit the telegram, we use an average electric power P during a certain time τ, and no heat. The entropy of the cable is not modified at first, but if we do not receive the message, and leave it on the cable (now short circuited at both ends), it will propagate back and forth and be progressively attenuated by Joule effect in the resistances. Finally the whole electric energy $P\tau$ is changed into heat. This final stage corresponds to an entropy S larger than the initial entropy S_0.

We can explain this situation by saying that the transmission of signals initially introduced on the cable a negentropy

$$\Delta N = S - S_0 \tag{5}$$

which was later on dissipated and lost in the resistances. The negentropy ΔN is the quantity which we want to compute. Some additional specifications are required at this point.

We may first assume a rather *strong coupling* between the signals and the physical system (the cable). This is Raymond's assumption "that the heat capacity of the closed system is so large that the change in temperature with operation of the communication device is negligible." This leads to a final entropy

$$S = S_0 + \frac{P\tau}{T}; \tag{6}$$

hence

$$\Delta N = \frac{P\tau}{T} = k\nu_M \frac{P}{P_n}\tau \tag{7}$$

with the help of Eq. (3).

The negentropy ΔN represents information flowing at the rate of $\Delta N/\tau$ per second, hence the channel capacity (per second)

$$C_1 = \frac{\Delta N}{\tau} = k\nu_M \frac{P}{P_n}. \tag{8}$$

Let us compare the result with Shannon's formula (4). We see that both formulas agree only for very small values of the ratio P/P_n when the logarithm can be expanded and the expansion reduced to its first term only.

Shannon's formula was obtained under special assumptions which actually do not correspond to the model just discussed. Shannon assumed a loss less cable, with no ohmic resistance. This means *no coupling* between the electric currents flowing on the cable and the physical cable itself. The source of noise is supposed to be outside the cable and to act independently. There is no direct mechanism here to produce the final increase of entropy that was found in our previous example. We may however keep the essential idea of no coupling between signals and cable and assume a dissipation mechanism of a different nature. The increase in entropy might be due to signal distortion, or to irregular changes in the spacing of current pulses and in the intensity of the pulses.

The initial state was an organized system of pulses carrying the desired information, and the final state is a disorganized and disordered succession of pulses, distributed at random, and carrying no intelligible information. Our system has n oscillating degrees of freedom[7] and a random distribution of energy E_r corresponds to a temperature T.

$$\begin{aligned} n &= \nu_M \tau \\ E_r &= nkT. \end{aligned} \tag{9}$$

If we now assume the random energy E_r to be progressively increased by steps dE_r, the entropy increase is

$$S - S_0 = \int_{E_1}^{E_2} \frac{dE_r}{T} = nk \int_{E_1}^{E_2} \frac{dE_r}{E_r} = nk \ln \frac{E_2}{E_1}. \tag{10}$$

The initial state E_1 corresponds to noise power P_n only, (Eq. 3) running for a time τ. The final energy level E_2 is obtained after all the additional energy $P\tau$ has been distributed at random between the n oscillators, and added to noise power

$$E_1 = P_n\tau \quad E_2 = (P_n + P)\tau; \tag{11}$$

hence the negentropy

$$\Delta N = S - S_0 = nk \ln \frac{P_n + P}{P_n}.$$

This yields a channel capacity (per unit time)

$$C_2 = \frac{\Delta N}{\tau} = \frac{n}{\tau} k \ln\left(1 + \frac{P}{P_n}\right) = k\nu_M \ln\left(1 + \frac{P}{P_n}\right) \tag{12}$$

which is identical with Shannon's formula (4). The difference between our previous result (8) and the new one (12) comes from the fact that our original model maintained a constant temperature T, while the new model exhibits a raise in temperature proportional to the raise in random energy E_r with Eq. (9) valid at all times. This raise in temperature decreases the entropy difference (10) and results in a lower capacity for the channel.

[7] It is a well-known result of signal analysis that a system of signals with frequencies $\nu \leq \nu_M$ and a duration τ has n degrees of freedom. We can check it by comparison with Eq. (3)

$$nkT = E_r = P_n\tau = kT\nu_M\tau.$$

The logarithm appears here is a similar fashion to the well-known $\ln V$ term in the entropy of an ideal gas.

In conclusion we must emphasize the physical meaning of this discussion: Shannon's formula really represents a limit that cannot be overcome since it corresponds to optimum conditions with maximum efficiency unity in the transformation of negentropy into information. This results from the fact that the computations based on information (Shannon) or on negentropy, give identical results, provided all quantities are measured in the same unit system.

If information is measured in bits and entropy in thermodynamical units, one finds that each bit corresponds to $k \ln 2 \approx 10^{-16}$ in thermodynamical units, since this is the ratio of units. This proves again that a bit of information can never be obtained for less than $k \ln 2$ in entropy cost, as was proved in a previous paper.[1]

4. THE TRANSMISSION OF INFORMATION, LIVE INFORMATION

The problem discussed in Sec. 3 is restricted to the simplest case of a communication from *person to person*, either by telegraph or by some telephonic device. This is a well defined and very important problem, the results of which can be extended to a variety of similar situations; a person directly talking to another person, for instance.

The theory however is not necessarily restricted to this special case and can be extended to much more general problems, provided special attention is paid to the conditions of operation. We intend to discuss these general problems, and to show that our *negentropy principle of information* enables us to investigate all possible circumstances of communication of information and to draw some important conclusions about their fundamental laws.

The first interesting problem is *broadcasting*. Here we have one person talking to a whole audience: A—One speaker talking before a microphone connected to N cables with N different receivers. B—One given telegram sent to N different stations. C—A speaker in front of an audience in an auditorium. D—Radio broadcasts. In cases A and B we obviously have examples of our previous "person to person" communication with N channels operated in parallel. The transmitter must provide the power required for all the channels. If the channels are identical, we simply need N times more power for the transmission. The cases C and D are less favorable, because of the large amount of power spread around and not utilized; sound absorbed in walls, furniture and other sound absorbing materials, in case C, radio waves lost and dissipated in space, in case D. Except for these very large additional losses, these last cases are similar to the first ones A, B. Altogether, we may reduce these problems to examples of single channels operated in parallel. All these problems have some common characteristics; power and negentropy are propagated together with the information. The receiver absorbs the energy and the negentropy in the same operation by which information is received. We shall characterize this situation by the expression *"live information."* The information is transmitted with the energy required for its detection. In the process of transmission, some energy is degraded, some negentropy is lost and information is lost by the same mechanism. The generalized Carnot principle states that negentropy must always decrease, and information follows the same law, in all these examples of live information.

5. THE PROBLEM OF WRITING AND READING

We find a completely different situation when we examine another type of transmission of information: writing and reading. There is obviously a certain expense of energy in the process of writing, but we do not find any trace of this energy on the written paper. There is no visible negentropy left either, and the information is there, completely dissociated from these other elements. This is not an isolated example, since a similar situation is found in all cases of recorded information: printed matter, phonograph records, magnetic tape, punched tape and punched cards, blueprints, etc. In these examples we have information stored and no energy nor negentropy connected with it. Furthermore, we can read out the information without destroying it. This seems to account for an unlimited possibility of multiplying information.

We had a principle by which information and negentropy were always decreasing. Here we have information without negentropy, and an apparent possibility for increasing indefinitely the information.

Let us specify the problem on a typical example. An author writes a book and the book is published. We do not intend to discuss the problem of thinking and the work of the author. This is the human element which we have been very cautious never to include in our discussions. It does not make any difference whether we consider a book written by an author, or a mere collection of numerical data: stock exchange quotations, weather observations, etc. If the manuscript is N letters long, it contains a certain amount of information. Our definitions do not go beyond that point.

The book is printed to 1000 copies, and each copy is read by 100 persons. The total amount of information has been multiplied by a factor 10^5. Instead of decreasing, information has been increasing, and we cannot see any limit to this increase.

Closer observation reveals that we shall have the usual causes for dissipation: misprints, errors of transmission, lack of understanding, etc. The actual factor will be less than 10^5, but it can still be very large.

The problem requires a closer examination in order to show how the negentropy principle of information can be applied to these situations.

6. DEAD INFORMATION AND HOW TO BRING IT BACK TO LIFE

We characterize the preceding situations as cases of stored dead information, with no energy and no negentropy connected with it. The point is that in order to read out the information we must re-energize it. An additional source of energy is absolutely necessary for the reading, and this source of energy provides the negentropy which is changed into new information.

We need a source of light to read a book. The phonograph does not work without a motor to turn the records.

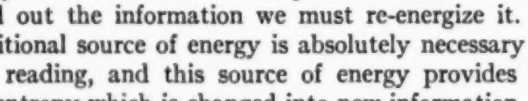


FIG. 1.

A blind person reading a book printed in the Braille alphabet has to move his finger on the paper and this motion requires work. We need light and photocells to read punched tape, or we may use a battery and brushes passing current through the holes of a punched card. In all these examples, the amount of negentropy taken from the external source is larger than the information obtained. The case of reading a punched tape is a typical example, on which we can compute both entropy and information, and prove the relation between them.

Information is recorded on punched tape in binary digits (Fig. 1). The zeros are represented by holes and the ones by unperforated tape sections. Each section of the tape contains one bit of information, or

$$\Delta I = k \ln 2 \quad \text{per section} \tag{13}$$

in thermodynamical units, when k is Boltzmann's constant. Our reading device (Fig. 2) uses a source of

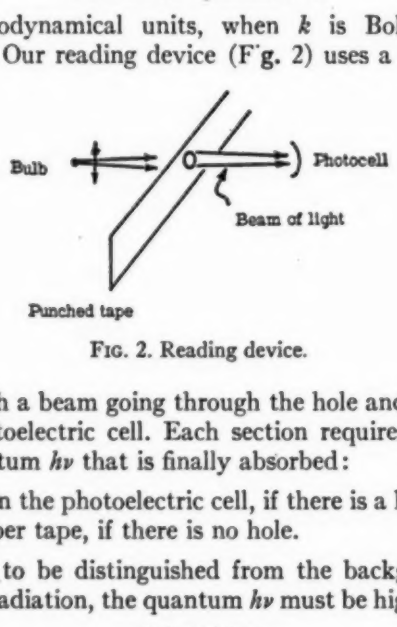


FIG. 2. Reading device.

light, with a beam going through the hole and received in a photoelectric cell. Each section requires at least one quantum $h\nu$ that is finally absorbed:

either in the photoelectric cell, if there is a hole, or in the paper tape, if there is no hole.

In order to be distinguished from the background of thermal radiation, the quantum $h\nu$ must be high enough

$$h\nu \gtrsim kT \ln 2 \tag{14}$$

as was proved in a preceding paper.[1] When the quantum is absorbed, there is a loss in negentropy (an increase ΔS in entropy).

$$\Delta S = \frac{h\nu}{T} \gtrsim k \ln 2 \tag{15}$$

and we have the relation

$$\Delta(S - I) \gtrsim 0 \tag{16}$$

which shows that our generalized Carnot principle is satisfied. In this simple case we prove that the information comes from the negentropy of the light beam, and we can theoretically reach the limit corresponding to an efficiency unity. The problem escaped notice because of the very small quantity of negentropy required

$$k \ln 2 \approx 10^{-16} \quad \text{in cgs centigrade degrees.} \tag{17}$$

The preceding type of punching is not too reliable, since the "one's" are not really recorded. A plain piece of tape, with no holes would be read as a string of "ones," which is wrong. A safer method would be to use two different types of holes, on two different levels (Fig. 3) punched in a reflecting metallic tape. This would require just as much energy for the reading operation, hence a ΔS just as large as in the preceding device.

FIG. 3.

We may use a system of brushes and a battery passing a current through the holes. The current is read on an ammeter or recorded in a relay. According to our discussion in the preceding paper,[1] the limit in this case is again $k \ln 2$ per reading. In a mechanical device, like a phonograph record, small positive or negative pulses can be read out only if the mechanical system yields an energy larger than $kT \ln 2$ per pulse, hence an entropy increase $k \ln 2$. Similar remarks apply to pulses recorded on magnetic tape. In every case, the negentropy lost in the reading operation is larger than the information obtained and the generalized Carnot principle is always satisfied.

Similar remarks apply to the problem of information recorded in blueprints, that can be used to build special machines. Each time we build a machine we must read out the blueprints. The Reading operation absorbs negentropy, part of which is changed into information imbedded in the machine built according to blueprints.

7. WRITING AND PRINTING

We can now consider another problem: how much negentropy is spent in the process of writing? This negentropy is lost, but we obtain a new copy of the information, hence the printing mechanism is another example of the transformation of negentropy into information.

Photographic printing is the simplest example to discuss: we put the punched tape on a photographic paper and send a beam of light on each position. Each time, we have to use one quantum $h\nu$ or more, in order

to print the position of a hole. The situation is similar to the case of Sec. 6 and leads to similar conclusions: printing a new copy costs exactly as much as reading it. The new information is paid for in negentropy.

In all these discussions, we ignored typical irreversible processes: developing the photographic paper, or amplifying the electric impulses, etc. All these refinements of the technique cost a great deal in entropy, and the total cost of a reading or printing operation will be very much higher than the limit we computed.

The essential point is: every information is paid for in negentropy. The amount of negentropy required is finite, but it is so small (10^{-16} per bit) that it has been completely ignored up to now. The connection between entropy and information is absolutely needed for consistency. The problems discussed in this paper would appear helpless otherwise. The very small value 10^{-16} of entropy required per bit of information plays an essential role in modern life, and makes it possible to communicate information at a negligible cost.

Another important remark is the law of *decay*: any system on which information has been recorded will deteriorate in the long run. Furthermore the information itself may progressively lose its value. For both reasons, recorded information follows a general law of decay, which is consistent with the generalized Carnot principle. The decay will, however, be much slower for dead (recorded) information than for live information.

8. INFORMATION AND ORGANIZATION

We have been able to analyze the information content of a written sentence, of a set of numbers, and of many other situations where our statistical definition of information could be directly applied.[8] In a similar way, we may compute the information contained in a blueprint, provided this blueprint can be explained with a finite number of symbols. Let us consider, for instance, the wiring diagram of some electric device; we may number from 0 to N all the terminals appearing on the diagram and explain the diagram by giving the numbers of the terminals connected together: a bracket (1023–0216) simply means that we must connect terminal 1023 to terminal 0216. For a finite number of terminals there is a finite number P of possible connections, out of which the wiring diagram selects one special system. It therefore contains an information (see paper 1)

$$I = K \ln P. \tag{18}$$

Let us again emphasize the exact meaning of such a definition: it does not make any difference between a wiring diagram corresponding to a working model and another diagram that would yield a worthless structure with shortcircuits or other defects. The information we compute is just the information given to the craftsman who is going to do the wiring, and has no specific idea of what the machine is going to do.

The question *why* this diagram has been selected and *how* it works is *not considered*. This was the problem of the scientist or the engineer who invented the device. It required thinking and scientific discussion, all brain operations which are beyond the scope of our present methods of investigation.

Let us take a specific example: the electric device contains n circuit elements, each of them with m terminals, and q terminals for outside connections (including ground or open connection). Altogether we have

$$N = nm + q \tag{19}$$

terminals, each of which can be connected to any other one, giving

$$P = N^N$$

different possibilities. The information contained in one wiring diagram is

$$I = K \ln P = KN \ln N. \tag{20}$$

Taking $K = \log_2 e \approx 1.4$, we obtain the information in *bits*

$$I = N \log_2 N.$$

Assuming reasonable orders of magnitude we consider the case

$$n = 999 \quad m = 10 \quad q = 10$$

hence

$$N = 10\,000 \approx 2^{13.3}$$

$$I = 133\,000 \text{ bits}. \tag{21}$$

This is a large figure in bits, but if we compute it in entropy units, we must multiply by 10^{-16} which is approximately the ratio of units and we obtain a negentropy

$$-S = 1.33 \quad 10^{-11} \tag{22}$$

which is completely negligible.

The machine built according to our specifications contains such a structural negentropy which represents the information or organization of the machine. This negative term however is completely negligible when compared with the total entropy of the machine. This example exhibits some characteristic features of the whole theory.

The negentropy principle of information requires that every piece of information be connected with a corresponding negative term in the physical entropy of a system. This connection is absolutely needed for consistency and was proved by the discussions of the preceding paper.[1]

Practically, these negative terms can be ignored in most problems because of the small coefficient (10^{-16}) introduced by the change from binary digits (bits) to thermodynamical units. The smallness of these terms is the fundamental reason why transmission of information by any practical method: writing, printing, telecommunication, is so inexpensive in entropy units, which also means inexpensive in dollar units. Modern life is based on these facts and would be completely different in a world where the negentropy of information would have a larger value.

[8] L. Brillouin, Am. Scientist 38, 596 (1950); J. Rothstein, Inst. Radio Engrs. conference (March, 1952).

Electrolytic Analog Transistor*

HARRY LETAW, JR., AND JOHN BARDEEN
Electrical Engineering Research Laboratory, University of Illinois, Urbana, Illinois
(Received September 8, 1953)

The electrolytic analog transistor is an operating model of the junction transistor which substitutes the reduced and oxidized forms of ions in solution for electrons and holes in a semiconductor. A base electrode makes a low-resistance contact to the solution and also serves to maintain the ratio of oxidized and reduced ions at an equilibrium value, thus establishing the potential of the solution. In *P-N-P* operation, a polarizable metal electrode (the emitter) is held at such a potential above the base electrode that the majority ions, the reduced form, are readily oxidized at its surface. An identical electrode, the collector, is placed within a few tenths of a millimeter of the emitter. It is biased in such a way that the minority ions (the oxidized form) which diffuse from the emitter to its surface are reduced. Utilizing several oxidation-reduction couples and different base electrodes, it has been found that values of $\alpha = -(\partial I_c/\partial I_e)_{V_c}$ in the neighborhood of unity can be obtained. Because of the relatively low mobilities of ions in solution, the device operates only at low frequencies. The device may also have some application as a constant current element in transistor bias circuits.

I. INTRODUCTION

THE operation of both point-contact[1] and junction[2] transistors depends on current flow in semiconductors when carriers of two signs, electrons and holes, are present. For example, in a *P-N-P* junction transistor, holes injected at one *P-N* junction, the emitter, diffuse across the narrow *N* region to the collector junction. The magnitude of the current depends upon the voltage difference between the emitter and a base electrode which makes a low-resistance contact to the *N* region. When the collector is biased in the reverse direction, most of the current flowing to the collector comes from the emitter and the current is nearly independent of collector voltage.

We have constructed an electrolytic analog of the junction transistor in which the flow of ions in solution replaces the flow of electrons and holes in the semiconductor. An oxidation-reduction couple forms a suitable system.

An ion in a reduced state in a solution which is normally oxidized (consists largely of ions in the oxidized state under the chosen conditions) corresponds to an excess electron. An ion in an oxidized state in a normally reduced solution corresponds to an electron defect or hole. In the analog of the *P-N-P* transistor the electrolyte is normally in the reduced state. One electrode, the analog of the base, makes a low-resistance contact to the solution. The emitter and collector of the analog are similar highly polarizable electrodes such that little current flows unless the voltage difference between one of the electrodes (the emitter) and the base exceeds a critical value (the decomposition potential) for oxidation to occur. The collector is maintained at a lower voltage than the emitter and is placed very close to it. Ions oxidized at the emitter diffuse to the collector where the reverse reaction occurs with the return of the ions to the reduced state. If the solution is normally in the oxidized state, one has the analog of the *N-P-N* transistor. The polarities are reversed and the reaction occuring at the emitter is a reduction.

If the conductivity of the electrolyte is sufficiently high, the electric fields in solution are small and the ions move mainly by diffusion. Inert electrolytes may be added to increase conductivity.

The frequency response of a transistor is limited by the time taken for carriers to diffuse from the emitter to the collector. In germanium transistors, transit times are of the order of 10^{-7} second. Since the mobilities of electrons and holes in germanium are at least 10^7 times larger than those of ions in solution, the transit time in the electrolytic transistor is one second or longer. Transient phenomena which occur in the megacycle range in ordinary transistors occur in the analog on a greatly expanded time scale and may thus be studied more readily.

It is believed that the analog may be useful in studying flow problems in transistor structures. The geometrical configuration of electrodes in solution can be changed readily to test a variety of structures. It is not yet certain whether the analog can be made sufficiently close so that it can be used in a quantitative as well as a qualitative manner.

It is also believed that the electrolytic transistor may have useful applications of its own for current control devices for which ordinary transistors are not suitable.

II. DESIGN OF THE CELL

In order that transistor action be observed, it is necessary that ionic or molecular species produced at one electrode control the current at the other, either through specific participation in the electrode process or else in the rate-determining step. To assure that such action occur in a reasonable time, and also to avoid losses through convection or off-axis diffusion, the emitter and collector must be quite close to one another. In the arrangement which we have used, the emitter

* Supported in part by the U. S. Office of Naval Research.
[1] J. Bardeen and W. H. Brattain, Phys. Rev. 74, 230 (1948).
[2] W. Shockley, Bell System Tech. J. 28, 435 (1949).

and collector are circular platinum electrodes 0.5 in. in diameter, one being moved by a micrometer screw arrangement and the other being fixed. The reverse sides of these electrodes are insulated from the solution.

The base electrode must serve to establish the potential of the solution. The most suitable electrode for this purpose is a relatively unpolarizable one which interacts reversibly with one or more of the ionic species in solution. For example, a mercury pool in contact with mercurous chloride and chloride ion is a satisfactory electrode. The base electrode serves to maintain the desired equilibrium between oxidized and reduced species.

The cell itself is an open framework constructed of lucite. All joints which come into contact with the solution as well as all submerged metallic parts, except for the emitter and collector faces, are thoroughly coated with a suitable cement. As is shown in Fig. 1, the electrical leads are insulated from the solution by being inserted through the cell framework. A piston-cylinder arrangement affords rather high orientational stability of the movable electrode with respect to the fixed one.

III. EXPERIMENTAL

The circuitry involved in the experimental arrangement was quite standard. Storage batteries were connected through potentiometer rheostats to the emitter and base and collector and base, respectively. Voltmeters and ammeters were connected in the usual way. This, along with the current and voltage convention for *P-N-P* operation, is shown in Fig. 2.

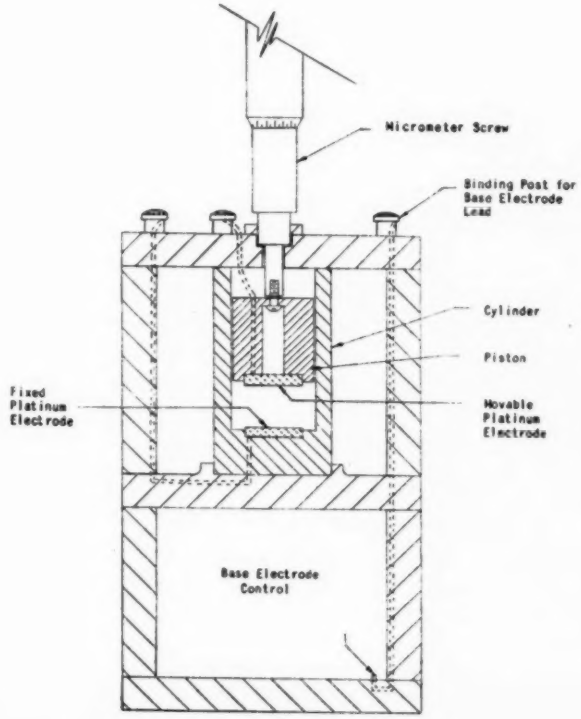


FIG. 1. Variable spaced electrolytic analog transistor cell.

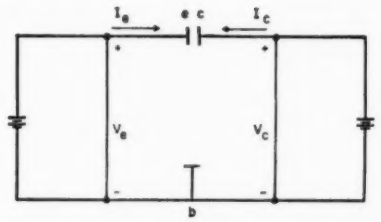


FIG. 2. Circuit schematic and current-voltage conventions for the *P-N-P* electrolytic transistor.

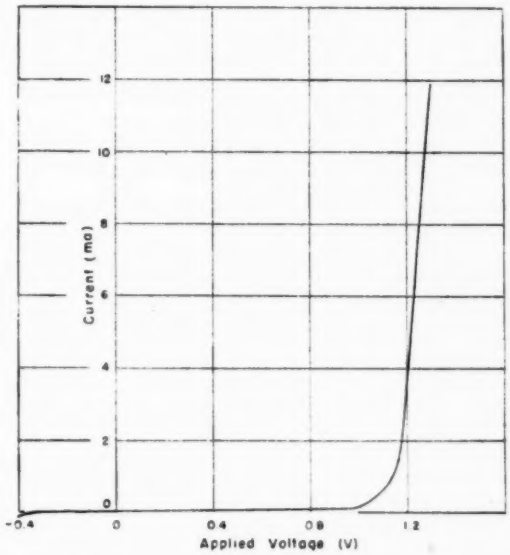


FIG. 3. Rectification characteristic for the chloride-chlorine couple.

The oxidation-reduction couples utilized in the present research had an important property in common, namely, that in contact with the base electrode the equilibrium concentration ratios of the two species yielded large numbers. It is easily shown that in one of the solutions used the equilibrium ferrous ion concentration is greater than that of ferric ion by a factor of 10^7.

In general, the techniques used were simple. A solution of the desired components was prepared and allowed to stand over the base electrode in a beaker. The cell was then placed into the beaker with the desired gross setting of the electrode spacing. No measurements were taken until stirring subsided. It was found that the device quickly recovered from ordinary vibrations in the laboratory. If the electrodes were spaced less than a few tenths of a millimeter apart, vibrations produced no observable effects. This spacing also eliminated interference which could result from thermal or density gradients in the solution by assuring that extraneous flow between the electrodes was essentially laminar and, if not of short duration, at least uniform.

TABLE I. Emitter and collector characteristics of the electrolytic transistor.

(Chloride-Chlorine Couple)			
I_e (ma)	V_e (v)	I_c (ma)	V_c (v)
0.154	1.00	−0.123	0.00
0.482	1.05	−0.418	0.00
2.42	1.10	−2.17	0.00
5.17	1.15	−4.80	0.00
7.32	1.18	−6.89	0.00
9.09	1.23	−8.62	0.00
0.0467	1.00	−0.0836	0.30
0.389	1.05	−0.337	0.30
2.21	1.10	−1.97	0.30
5.33	1.15	−5.02	0.30
7.43	1.20	−7.03	0.30
8.64	1.25	−8.23	0.30

A. Chloride-Chlorine Couple

A solution 1M with respect to NaCl and 0.2M with respect to HCl was prepared in distilled water. In two electrode operation, i.e., with the collector disconnected, the rectification curve shown in Fig. 3 was obtained. The base electrode was a mercury, mercurous chloride pool.

In transistor operation, V_c (collector voltage) was held constant while V_e (emitter voltage) was increased. From the data in Table I, it can be seen that the emitter was operated in the sharply rising portion of the voltage-current curve shown in Fig. 3. As V_e increased, I_e (emitter current) increased. This resulted in an increase in $-I_c$ (collector current). The emitter current is plotted against the collector current in Fig. 4. The slope of this curve is $\alpha=(-\partial I_c/\partial I_e)_{V_c}$. This is essentially a measure of the efficiency of the modulation of the current in the collector loop by the emitter. With this particular value of electrode spacing, $\alpha=0.96$. This corresponds to collection of 96 percent of the chlorine produced at the emitter. Since all of the data fall upon one curve, it is seen that α is essentially independent of the collector voltage in the range studied.

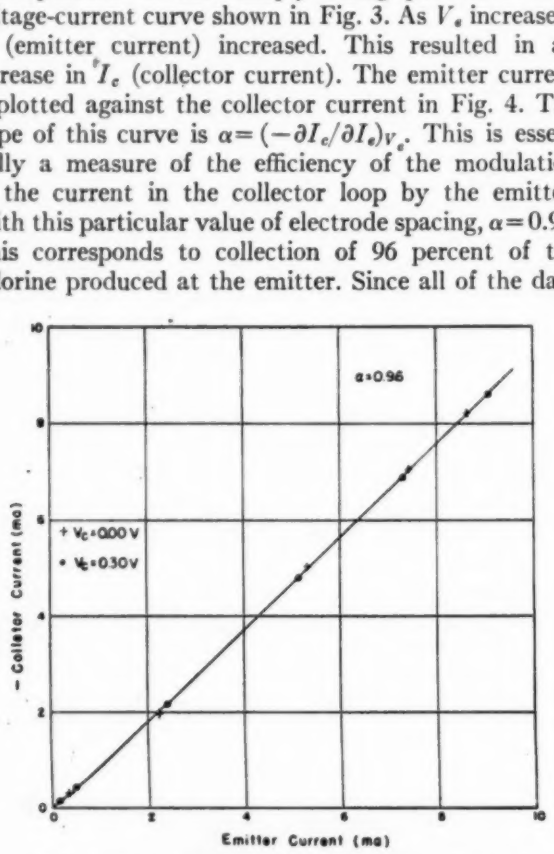


FIG. 4. Electrolytic transistor current characteristics for the chloride-chlorine couple.

If the process under investigation is diffusion limited, one would suspect that the modulation as reflected in α is a function of the emitter-collector spacing d. The parameter d was varied between 0.1 and 1.2 mm. These data are presented graphically in Fig. 5. It will be noted that α does decrease with increasing d. Figure 6 is a graph of d *versus* α.

The chloride-chlorine couple is more stable than that attributable to the oxy-acids of chlorine except for the chloride-perchlorate couple. The latter couple is eliminated because it is essentially irreversible at room temperature. Further, one observes the presence of Cl_2 in the solution surrounding the emitter by its characteristic yellow-green color.

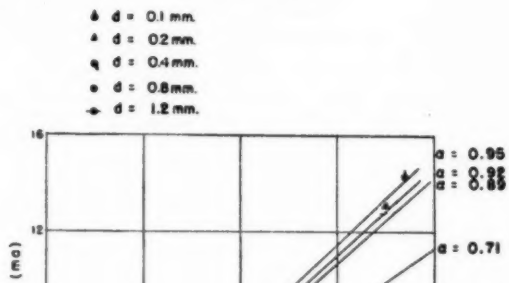


FIG. 5. Electrolytic transistor current characteristics for the chloride-chlorine couple as a function of emitter-collector spacing

B. Ferrous-Ferric Couple

A ferrous ion solution containing 0.25M $FeSO_4$ and 1.5M HCl was investigated. The rectification characteristic curve shown in Fig. 7 indicates the appearance of the expected limiting current. The electrode separation parameter d being about 0.08 mm, one expects α to be very nearly unity. It is shown graphically in Fig. 8 that this is the case. The corresponding data are given in Table II.

The collector current plotted against the collector voltage for this solution is shown in Fig. 9. The resultant output characteristic indicates that the collector operates independently of its bias in the region from −0.25 v to +0.25 v.

C. Bromide-Bromine Couple

It was pointed out to us by Professor H. A. Laitinen that the highly reversible bromide-bromine couple

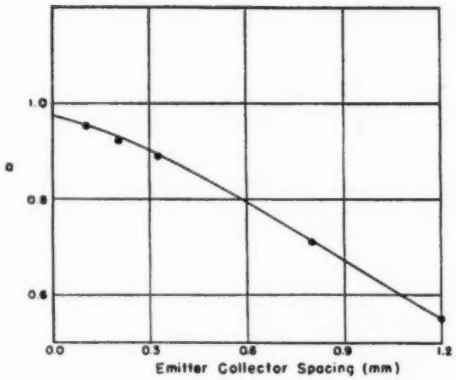


FIG. 6. α as a function of emitter-collector spacing for the chloride-chlorine couple.

might be satisfactory for use in this apparatus. This was attempted with the use of two different base electrodes, mercury, mercurous bromide and silver, silver bromide.

(1) *Mercury, Mercurous Bromide Electrode*

The rectification characteristic obtained using a solution of 1M NaBr was identical to that shown in Fig. 3 except for the fact that it was not quite so flat in the low-voltage region and the break occurred near 0.8 v.

It is well known that bromine decomposes water in solutions of low acidity. In the present case, this would result in an increase in α with a decrease in pH. This was investigated in four solutions of pH 7, 4, 2.5, and 1 with the volume and bromide ion concentration remaining constant. Decrease in pH was brought about through the addition of HBr. It was found that the value of α was constant at 0.98. Emitter and collector characteristics for pH 7 and pH 1 are reproduced in Table III.

It is not improbable that in a slightly basic solution this property could be utilized to simulate the phenomenon of recombination of injected carriers in the base region of a junction transistor.

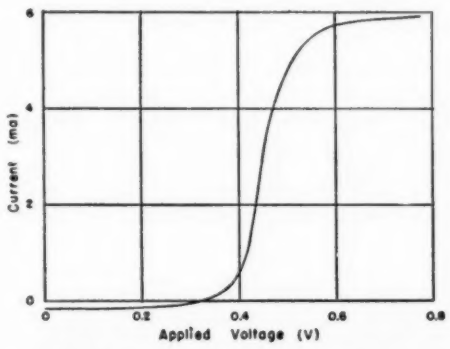


FIG. 7. Rectification characteristic for the ferrous-ferric couple.

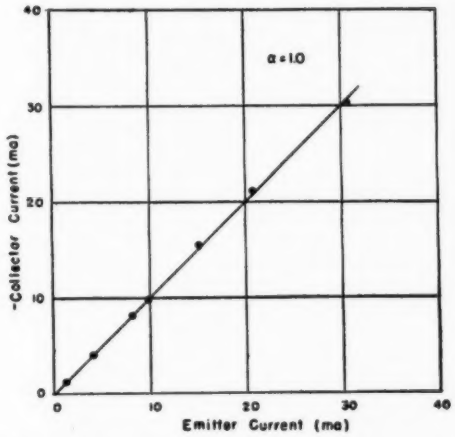


FIG. 8. Electrolytic transistor current characteristic for the ferrous-ferric couple.

(2) *Silver, Silver Bromide Base Electrode*

A slight modification of the base electrode contact was made for this experiment. The rectification characteristic using a solution 1M with respect to NaBr and 0.1M with respect to HBr was extremely flat from 0.0 to 0.9 v. Although no gas evolution was observed until the applied voltage was in the neighborhood of 2 v, there is a suggestion of an inflection point in the rectification curve at 1.1 v. This, if it is real, would indicate that a process other than the oxidation of bromide to bromine occurs near this voltage.

TABLE II. Emitter and collector characteristics of the electrolytic transistor.

(Ferrous-Ferric Couple)			
I_e (ma)	V_e (v)	I_c (ma)	V_c (v)
1.30	0.32	− 1.37	0.00
4.05	0.40	− 4.00	0.00
8.25	0.43	− 8.12	0.00
10.0	0.44	− 9.86	0.00
15.2	0.46	−15.5	0.00
20.8	0.50	−21.0	0.00
30.6	0.57	−30.3	0.00

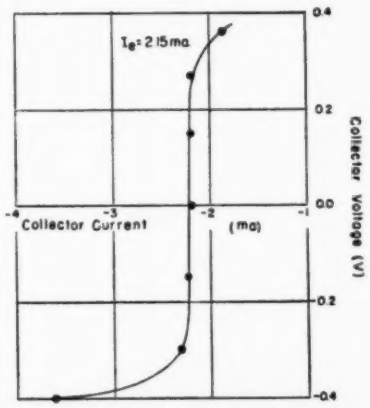


FIG. 9. Electrolytic transistor output characteristic for the ferrous-ferric couple.

TABLE III. Emitter and collector characteristics of bromide-bromine couple as a function of pH.

	I_e (ma)	V_e (v)	I_c (ma)	V_c (v)
pH 7	0.506	0.89	−0.480	0.25
	1.41	0.91	−1.32	0.25
	2.67	0.93	−2.57	0.25
	5.16	0.96	−5.05	0.25
	7.63	0.98	−7.52	0.25
	9.62	1.00	−9.50	0.25
pH 1	0.376	0.89	−0.363	0.25
	0.951	0.92	−0.920	0.25
	2.24	0.92	−2.12	0.25
	4.84	0.95	−4.67	0.25
	6.87	0.97	−6.70	0.25
	9.33	0.99	−9.10	0.25

In operation as a transistor, this solution yielded $\alpha=0.92$ with a spacing of about 0.04 mm. The output characteristic, shown graphically in Fig. 10, was exceptionally flat.

IV. THEORY

The conductivities of the solutions used are so high that the flow of ions between the emitter and collector is mainly by diffusion. Electric fields and consequently potential differences in the body of the electrolyte are small. There will, however, be space-charge double layers extending outward a very small distance from the electrodes. When voltages are applied to the electrodes, part of the voltage drop may be across the double layers and part may appear through the body of the solution as a concentration difference.

Let us first consider the diffusion of ions between the emitter and collector when the spacing is such that α is nearly equal to unity. Let c_1 and c_3 be the concentration of the majority species at the emitter and collector, respectively, and let c_2 be the concentration at the emitter of the (minority) species of ion produced there. These concentrations refer to positions just outside of the double layers at the electrodes. We shall assume that the collector is biased in such a way that the concentration of minority ions at the collector is practically zero. The variation of the concentrations between the emitter and collector is then as indicated in Fig. 11. In the ferrous-ferric couple c_1 and c_3 refer to the ferrous ion and c_2 to the ferric ion.

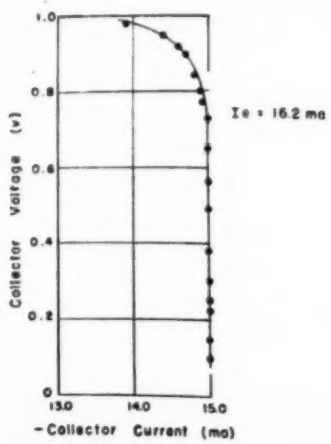


FIG. 10. Electrolytic transistor output characteristic for the bromide-bromine couple and silver-silver bromide base electrode.

Since under steady-state conditions the particle currents of the two species of ions must be equal, the current density may be expressed as

$$I=eD_1(c_3-c_1)/d=eD_2c_2/d, \tag{1}$$

where D_1 and D_2 are the diffusion coefficients and d the emitter-collector spacing. Equation (1) may be written

$$bc_2=c_3-c_1, \tag{2}$$

where $b=D_2/D_1$.

If it is assumed, as is likely the case under the conditions of our experiments, that the total concentration remains constant, we have

$$\tfrac{1}{2}c_2+\tfrac{1}{2}(c_1+c_3)=c_{10}. \tag{3}$$

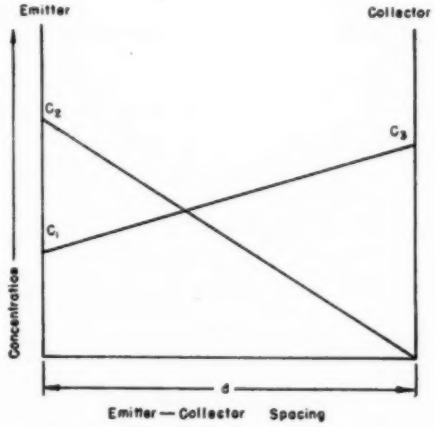


FIG. 11. Electrolytic transistor concentration convention.

where c_{10} is the initial equilibrium concentration of the majority species.

One further relation is required to determine how the concentrations vary with applied voltage. If it is assumed that the current is limited by diffusion, the voltage drop will occur in the body of the electrolyte and for a simple couple such as ferrous-ferric,

$$c_2/c_1=\exp[e(V-V_0)/kT], \tag{4}$$

where V is the emitter voltage and V_0 is the standard oxidation potential of the cell which is computed by the algebraic addition of the standard electrode potentials of the operating couples. If c_{10} and c_{20} are the equilibrium concentrations of the two species

$$c_{20}/c_{10}=\exp(-eV_0/kT),$$

Eq. (4) may be written

$$c_2/c_1=(c_{20}/c_{10})\exp(eV/kT). \tag{5}$$

Solution of Eqs. (2), (3), and (5) for c_2 gives

$$c_2=c_{10}\exp[e(V-V_0)/kT]/\{1+[(1+b)/2]\exp[e(V-V_0)/kT]\}. \quad (6)$$

If one defines

$$I_0=eD_2c_{10}/d, \quad (7)$$

the current density is

$$I=I_0\exp[e(V-V_0)kT]/\{1+[(1+b)/2]\exp[e(V-V_0)/kT]\}. \quad (8)$$

For this case, in which the current is limited by diffusion, the current should be inversely proportional to the electrode spacing d. If, on the other hand, the current is limited by an electrode reaction, it will be essentially independent of d. Figure 12 is a plot of I *versus* $1/d$ for the chloride-chlorine couple. When $1/d$ is large (d small), the current approaches a constant value which indicates that in this limit the current is determined by an electrode reaction. The decrease in current with d for the larger spacings indicates that the current then is diffusion limited. The ferrous-ferric couple was observed to behave similarly.

The theory for the chloride-chlorine couple is somewhat more complicated than that given above because the chlorine is in solution in molecular form. The equilibrium concentration $[Cl_2]_0$ of Cl_2 molecules in solution, determined from the reaction

$$Hg+\tfrac{1}{2}Cl_2\rightleftharpoons\tfrac{1}{2}Hg_2Cl_2,$$

is

$$[Cl_2]_0^{\frac{1}{2}}=\exp(-eV_0/kT), \quad (9)$$

where V_0 is determined from the difference in standard oxidation potentials for the half-reactions:

$$Cl^-\rightleftharpoons\tfrac{1}{2}Cl_2+e \qquad (E^0=-1.3583\text{ v})^3$$
$$Hg+Cl^-\rightleftharpoons\tfrac{1}{2}Hg_2Cl_2+e \qquad (E^0=-0.2676\text{ v})^3.$$

Thus,

$$V_0=1.3583-0.2676=1.0907\text{ v}. \quad (10)$$

If the current is diffusion limited,

$$[Cl_2]^{\frac{1}{2}}_{em}/[Cl^-]_{em}=([Cl_2]_0^{\frac{1}{2}}/[Cl^-]_0)\exp(eV/kT), \quad (11)$$

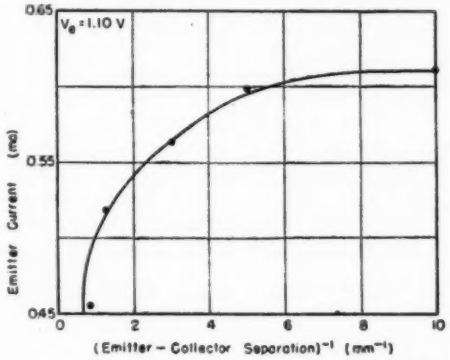


FIG. 12. Emitter current as a function of emitter-collector spacing.

[3] W. M. Latimer, *Oxidation Potentials* (Prentice-Hall, Inc., New York, 1938).

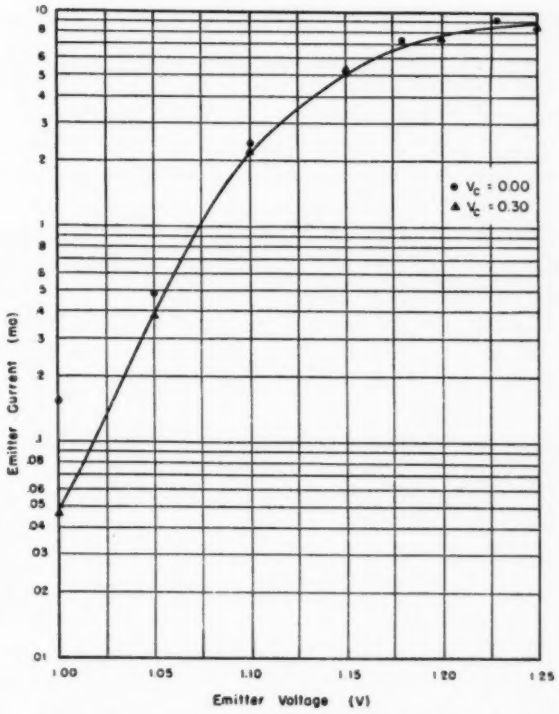


FIG. 13. Electrolytic transistor emitter voltage-log emitter current characteristic.

where the subscript "em" refers to the concentration just outside the double layer at the emitter. Equation (11) replaces Eqs. (4) and (5). Equations (1) and (3) also require modification. Let

$$c_2=[Cl_2]_{em}, \quad c_1=[Cl^-]_{em}.$$

Equation (1) becomes

$$I=eD_1(c_3-c_1)/d=2eD_2c_2/d, \quad (12)$$

since each Cl_2 molecule contains two atoms. Similarly, Eq. (3) is replaced by

$$c_2+\tfrac{1}{2}(c_1+c_3)=c_{10}. \quad (13)$$

Equations (11), (12), and (13) must then be solved for the concentration and current density. One is led to a quadratic equation for c_1 whose solution will not be given explicitly. When V is small, c_1 is nearly equal to the equilibrium value c_{10}, and c_2 (and thus I) varies as $\exp(2eV/kT)$. When V is very large, the current tends to saturate at a constant value.

If the current is limited by an electrode reaction involving the majority species, one expects

$$I\sim c_1\exp(\beta eV/kT), \quad (14)$$

where β is a numerical parameter which depends upon the nature of the reaction. This again leads to an exponential variation of current with voltage when V is small and saturation at a constant value when V is large. This behavior is shown for the chloride-chlorine case in Fig. 13.

V. DISCUSSION

The behavior of the electrolytic analog transistor under the conditions described above is basically as predicted. This extension of the phenomenon of transistor action to the liquid state confirms the general applicability of the physical analysis of the semiconductor transistor.

Two types of electrochemical reactions have been studied. These are the conversion of an ion to an uncharged molecule and the change in valency of an ionic couple. Metal-metal ion couples could be selected in various combinations as could metal, metal salt electrodes in the presence of their anions. The operation of the cell would also be possible with such organic couples as quinone-hydroquinone under various pH conditions. Nonaqueous solvents might also be used.

The thermodynamic basis of collector activity is easily understood by a consideration of the details of emitter action. The entire cell is initially in equilibrium with the establishment of a certain stoichiometric balance as dictated by the standard electrode potentials of the operating couples and the ambient temperature and pressure. Upon the insertion of an emf in the emitter circuit, the emitter and base tend toward another stoichiometric balance.

Consideration of a definite set of couples will facilitate the explanation of the action in the collector circuit. If the emitter and collector are reversible to the ferrous-ferric couple and the base is a calomel electrode (which is reversible to chloride ion), the introduction of an emf into the emitter circuit results in the production of ferric ions at the emitter and chloride ions at the base. Ferric ions which diffuse to the collector are reconverted to ferrous ions. Under steady-state conditions the current to the base electrode is small. Thus, the collector circuit operates to restore equilibrium even though no bias voltage is applied. If, in this case, a positive bias is applied, it should in no way affect the tendency of the collector to seek equilibrium unless it is of such magnitude as to be near or greater than the decomposition voltage of the ferrous-ferric couple.

An unexpected result of this investigation was the observation of the fact that the limiting current of several solutions as determined from rectification characteristics was enhanced when the device was placed in three electrode operation. The enhancement was a factor of about two for the chloride-chlorine couple and about eight for the ferrous-ferric couple.

At these spacings, the current is limited by diffusion of the majority ion in solution (in these cases, the chloride and ferrous ions) to the emitter. The width of the diffusion region is generally determined by convection currents in the liquid. If the spacing between the emitter and collector is less than this width, the diffusion path is decreased and the current enhanced. The concentration of the majority ion at the collector is determined by the potential of the collector relative to the solution and is maintained by the reverse reaction which occurs at the collector.

The graph shown in Fig. 10 demonstrates the behavior of the output characteristic under varying bias conditions. This property of the electrolytic analog transistor has led to a consideration of its utility as a constant current element in transistor biasing circuits. As was mentioned above, this device should not respond to alternating currents above one cycle or so.

This model of the electrolytic transistor is limited in current gain to unity in grounded base operation just as is the junction transistor. Since the output impedance is much higher than the input impedance, power gain can be obtained at very low frequencies.

Experiment indicates that the current is not strictly diffusion limited at small emitter-collector spacings. The rate of the reaction occurring at the electrode controls the process under these conditions. At larger distances, the steady-state process is controlled by diffusion. Analysis of several curves such as that of Fig. 13 indicates that the value of β in Eq. (14), which is obtained from the lower asymptote to the curves, is nearly unity for the halide-halogen couples. The value of this parameter is approximately 1.5 if small quantities of cerous and ceric sulfates are present in the chloride-chlorine solution. Since it is probable that these ions in some way catalyze the rate-determining electrode process, this observation would indicate that the theoretical value of β is indeed not attained because of a slow step at the electrode. Though the lower asymptotic current was not attained in the case of the ferrous-ferric couple, extrapolation of the curve to a value of β equal to unity appeared to be reasonable.

The apparatus used in these experiments included bright platinum electrodes. It is evident that the value found for β depends strongly upon the instantaneous condition of the electrode surfaces.

Research is currently being carried on for the purpose of investigating the electrolytic transistor as a circuit element and to determine its behavior as a function of temperature.

ACKNOWLEDGMENTS

We gratefully acknowledge the advice of Professor G. L. Clark, Professor H. A. Laitinen, and Professor Sherlock Swann, Jr., of the Department of Chemistry. Professor T. A. Murrell and Professor L. M. Slifkin, Dr. S. R. Morrison, and Mr. N. Holonyak, Jr., Mr. M. F. Millea, Mr. R. S. Sirrine, and Mr. R. Sun of this laboratory have contributed materially to the completion of this phase of the problem.

Diffusion in Bimetal Vapor-Solid Couples*

R. W. BALLUFFI AND L. L. SEIGLE
Atomic Energy Division, Sylvania Electric Products Inc., Bayside, New York
(Received September 8, 1953; revised manuscript received December 28, 1953)

The intrinsic diffusivities of individual components in several binary systems have been determined with vapor-solid type diffusion couples. In this type of couple a high vapor-pressure component is diffused from the vapor phase into a slab of lower vapor pressure initially containing inert markers on its surface. Diffusivities may then be calculated from marker movement and diffusion-penetration curve data. Equations necessary for analysis of such couples are developed.

The alpha brass, copper-nickel, and silver-gold systems were investigated. Observations of polygonization during diffusion and dimensional changes normal to the diffusion direction were also made. Porosity formation, which occurred only in the vapor-solid silver-gold couples, was investigated in detail and was compared to porosity formation in corresponding sandwich couples. The general use and limitations of such couples in diffusion studies are discussed.

I. INTRODUCTION

DIFFUSION phenomena in metal systems may be conveniently studied in certain cases by the use of vapor-solid couples in which one component is transported to the specimen surface through the vapor phase.[1-3] In this type of couple, inert markers are initially placed on the surface of a solid slab which is then exposed to the vapor of a high vapor pressure component at diffusion temperatures. In general, the high vapor pressure component diffuses more rapidly, and as a result of the rapid inward diffusion from the vapor the slab expands and the markers are pushed outward with the expanding specimen. At the same time the markers become buried because of the smaller diffusional flow of the lower vapor-pressure component in the opposite direction.

A photomicrograph of the buried markers in a typical alpha brass couple of this type is given in Fig. 1, and a schematic diagram of the diffusion zone is shown in Fig. 2. Consideration of the diffusion flow of zinc into the specimen from the vapor and of copper in the opposite direction shows that the entire hatched area, including the cross-hatched area, is proportional to the amount of zinc diffused inward past the markers and that the smaller cross-hatched area is proportional to the amount of copper diffused outwards past the markers. The ratio of diffusivities of copper and zinc is given by the ratio of these two areas. In addition, the intrinsic diffusivities may be determined from marker movement and diffusion-penetration data.

Such couples possess a number of practical advantages. No welding or plating is required, and single crystals may therefore be conveniently handled. Inert markers are easily attached to the specimen surface, and their movements can be determined with particular ease and accuracy, since no graphical integration or placement of reference foils are necessary.[4] In addition, porosity formation is avoided in a number of systems where it occurs in sandwich-type couples. The chief limitation of the method is that only components of widely differing vapor pressure may be used. In order to obtain absolute values of the intrinsic diffusion coefficients the vapor pressure of the volatile constituent must also be high enough so that the concentration at the vapor-solid interface is maintained constant during diffusion. If this boundary condition is not met, absolute values of the diffusivities cannot be derived from the diffusion curves. However, the difference between intrinsic diffusivities may still be obtained from marker movement data.

II. MATHEMATICAL ANALYSIS

A proper analytical treatment of vapor-solid diffusion couples must take into account both the variation of diffusivity with concentration and dimensional changes caused by the absorption of mass from the vapor phase.

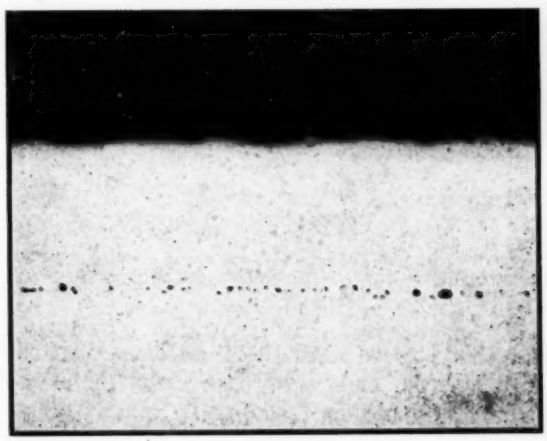


FIG. 1. Inert markers buried in alpha brass couple diffused 35.25 hr at 890°C. 1000X.

* This work was sponsored by the U. S. Atomic Energy Commission and was done under Contract AT-30-1GEN-367.

[1] R. W. Balluffi and B. H. Alexander, J. Metals 4, 1315 (1952).
[2] R. W. Balluffi and B. H. Alexander, J. Appl. Phys. 23, 953 (1952).
[3] R. W. Balluffi and B. H. Alexander, J. Appl. Phys. 23, 1237 (1952).
[4] L. C. C. da Silva and R. F. Mehl, Trans. Am. Inst. Mining Engrs. 191, 155 (1951).

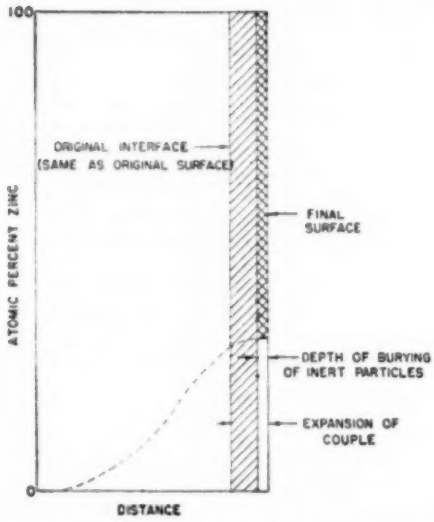


FIG. 2. Schematic diagram of alpha brass vapor-solid couple containing inert markers.

In order to obtain a mathematical analysis for the one dimensional case, it is assumed with Darken[5] that the diffusion process may be described by the use of individual diffusion coefficients for each element. If diffusion is considered as motion relative to inert markers then the total rate of transport of component 1 across any plane stationary with respect to the vapor-solid interface is

$$J_1=-D_1\frac{\partial C_1}{\partial x}+C_1v(x,t)+C_1\phi(t), \tag{1}$$

where

$$-\phi(t)=\left[D_1\frac{\partial C_1}{\partial x}\right]_{x=0}\frac{1}{C}. \tag{2}$$

In the above, v is the velocity of mass flow due to possible inequalities in the intrinsic diffusion coefficients, and ϕ is the velocity of mass flow arising from absorption of matter from the vapor.

Following Darken,[5] therefore,

$$\frac{\partial}{\partial x}\left[D_1\frac{\partial C_1}{\partial x}+D_2\frac{\partial C_2}{\partial x}-Cv(x,t)-C\phi(t)\right]=0, \tag{3}$$

and on integration

$$D_1\frac{\partial C_1}{\partial x}+D_2\frac{\partial C_2}{\partial x}-Cv(x,t)-C\phi(t)=I. \tag{4}$$

For a semi-infinite solid, as x becomes very large, $\partial C_1/\partial x$, $\partial C_2/\partial x$, and v approach zero; therefore the constant of integration, I, $=-C\phi(t)$, and

$$v=\frac{1}{C}\left[D_1\frac{\partial C_1}{\partial x}+D_2\frac{\partial C_2}{\partial x}\right]. \tag{5}$$

[5] L. S. Darken, Trans. Am. Inst. Mining Met. Engrs. **175**, 184 (1948).

Inserting the values for v and ϕ in Eq. (1), and expressing concentrations in terms of atom fractions

$$\frac{\partial N_1}{\partial t}=\frac{\partial}{\partial x}\left[(D_1N_2+D_2N_1)\frac{\partial N_1}{\partial x}\right]+\left[D_1\frac{\partial N_1}{\partial x}\right]_{x=0}\cdot\frac{\partial N_1}{\partial x}. \tag{6}$$

Trying for the solution $N_1=f(x/t^{\frac{1}{2}})\equiv f(\lambda)$

$$-\tfrac{1}{2}\lambda\frac{dN_1}{d\lambda}=\frac{d}{d\lambda}\left[(D_1N_2+N_2D_1)\frac{dN_1}{d\lambda}\right]+\left[D_1\frac{dN_1}{d\lambda}\right]_{\lambda=0}\cdot\frac{dN_1}{d\lambda}. \tag{7}$$

Therefore, $N_1=f(\lambda)$ is a solution when the origin is placed at the vapor-solid interface, providing the concentration at this plane remains constant. When this condition is satisfied the term $[D_1(dN_1/d\lambda)]_{\lambda=0}$ is a constant $\epsilon/2$ and Eq. (7) may be written

$$-\tfrac{1}{2}(\lambda+\epsilon)\frac{dN_1}{d\lambda}=\frac{d}{d\lambda}\left[(D_1N_2+D_2N_1)\frac{dN_1}{d\lambda}\right]. \tag{8}$$

By making use of the translation $\lambda'=\lambda+\epsilon$ in Eq.(8), the differential equation may be put into the familiar form,

$$-\tfrac{1}{2}\lambda'\frac{dN_1}{d\lambda'}=\frac{d}{d\lambda'}\left[(D_1N_2+D_2N_1)\frac{dN_1}{d\lambda'}\right], \tag{9}$$

where λ' is referred to an axis fixed at the undiffused end of the specimen. Inspection shows that

$$D=D_1N_2+D_2N_1, \tag{10}$$

where D is identical to the chemical diffusivity which is obtained from the usual Matano analysis of sandwich-type couples. The Matano analysis in its usual form, therefore, may be applied to the vapor-solid couple.†

The above relationships are valid for unidimensional diffusion, where dimensional changes normal to the direction of diffusion are negligible, no porosity is present, and lattice parameter changes are compensated for. Allowance is made for changes in lattice parameter by using a distance coordinate x in the diffusion direction, which is chosen so that equal increments of x correspond to equal numbers of atoms, and is defined by

$$dx=\frac{V_r}{V}dy, \tag{11}$$

where y is the experimentally measured coordinate, V is the local atomic volume, and V_r is the atomic volume

† The writers are indebted to Mr. H. Fara of Sylvania Electric Products Inc. for the above proof.

of a reference alloy.[6] In the systems investigated, the lower vapor-pressure material was always selected as the reference alloy.

An important consequence of the fact that $N_1=f(\lambda)$ satisfies the diffusion equation is that the concentration in the vicinity of the marker originally at the join (or originally at the surface of the vapor-solid couple) must be independent of time. This may be shown as follows‡

The velocity of any marker referred to the nondiffused end is given by:[5]

$$v_m=(D_1-D_2)\left(\frac{\partial N_1}{\partial x}\right)_m. \tag{12}$$

The diffusivities are functions of concentration alone and therefore

$$(D_1-D_2)=g\left(\frac{x_m}{t^{\frac{1}{2}}}\right).$$

Hence,

$$v_m=\frac{dx_m}{dt}=g\left(\frac{x_m}{t^{\frac{1}{2}}}\right)\cdot f'\left(\frac{x_m}{t^{\frac{1}{2}}}\right)\frac{1}{t^{\frac{1}{2}}}. \tag{13}$$

A general solution of Eq. 13 cannot be obtained since the functions f' and g are not known. However, by substitution in Eq. (13) it is seen that $x_m=Kt^{\frac{1}{2}}$ is a particular solution when K is chosen, so that

$$K=2g(K)\cdot f'(K). \tag{14}$$

Furthermore this is the particular solution for the marker initially at the join, since $x=0$ when $t=0$. Since the distance traveled by the marker and all composition levels varies as $t^{\frac{1}{2}}$, and initially the composition at the join is indeterminate, there must be one composition level which remains associated with the marker as it moves away from the original interface.

It follows from the preceding that the intrinsic diffusion coefficients determined from marker location refer to the particular composition at the marker, which remains constant during the experiment.

III. EXPERIMENTAL

Diffusion runs were made using the vapor-solid couples zinc-alpha brass, copper-nickel, and silver-gold. A major objective of the present work was to check the range of conditions under which the assumptions used in the preceding analysis are realized in vapor-solid diffusion couple experiments. The results for each system are presented separately below.

a. Zinc-Alpha Brass

Experiments designed to obtain the ratio of diffusivities of zinc and copper in alpha brass using vapor-solid couples have been previously described.[1] The ratio of diffusivities was determined by finding the ratio of the number of zinc atoms and copper atoms which diffused past the markers. When the inert markers remain at constant concentration, the ratio of diffusivities at the specific marker concentration is given by the ratio of the total number of atoms of each component diffused past the markers in opposite directions and, therefore,

$$D_1/D_2=Q_1/Q_2, \tag{15}$$

where Q_1 and Q_2, the total number of atoms of 1 and 2 diffused past the markers, may be obtained by suitable graphical integration under the diffusion curve.

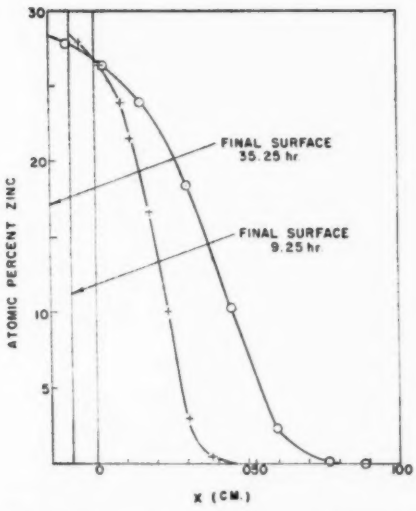


FIG. 3. Diffusion penetration curves of alpha brass couples diffused 9.25 and 32.25 hr at 890°C.

Two 0.25 in. thick −1.25 in. diameter discs of OFHC copper were prepared, having a coarse grain size of 1 mm. After inert particles of alumina were placed on the specimen surfaces the disks were sealed in evacuated quartz capsules along with a large zinc source of fine 29 atomic percent zinc-alpha brass chips, and were annealed at 890°C for 9.25 and 35.25 hr, respectively. The penetration curves are given in Fig. 3. The surface concentrations at both times were identical and were essentially the same as the composition of the chips after diffusion, indicating that approximate equilibrium was maintained between the chips and the surface. (The vapor pressure of zinc is about 50 mm Hg at this temperature.)

The expansions of the specimens and the marker movements, which were obtained from the difference between the specimen expansion and the depth of burying of particles (see Fig. 1), varied parabolically with time as shown in Fig. 4. Such a variation is in accordance with the solution $N_1=f(\lambda)$, which satisfies the differential equation for diffusion, providing the concentration at the vapor-solid interface is independent of time as in the present case. Satisfaction of the prescribed boundary conditions is also indicated by con-

[6] C. Wagner, J. Metals 4, 91 (1952).

‡ The discussion of this point by Darken (see reference 5) has been questioned by LeClaire (see reference 7), and it therefore seemed desirable to present a separate proof.

[7] A. D. LeClaire, Discussion to reference 5, Trans Am. Inst. Mining Met. Engrs. 175, 194 (1948).

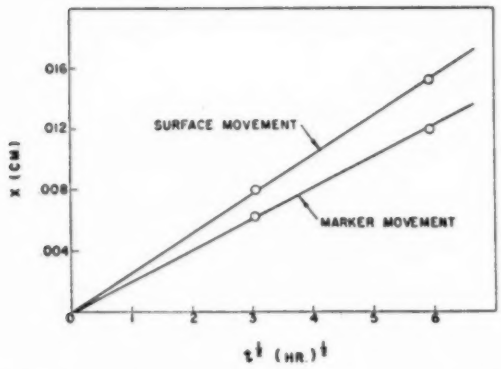


FIG. 4. Surface movement and marker movement *versus* $t^{\frac{1}{2}}$ for alpha brass couples at 890°C.

stancy of composition of the original interface ($x=0$ in Fig. 3), in conformity with the solution $N_1=f(\lambda)$.

Since the specimen expansion is solely the result of the gain in zinc, the total movement of the surface is quantitatively equal to the area under the diffusion curve. The position of the original interface was therefore obtained by graphical integration under the diffusion curves and was found to agree closely with the position found from the measured specimen expansion in each case. Careful metallographic examination of the diffusion zone revealed no evidence of porosity formation.

The chemical diffusivity was obtained as a function of concentration from the 35.25-hr diffusion curve by the Matano analysis and is shown in Fig. 5. These values agree fairly well with previous data obtained from sandwich-type couples, recently summarized by da Silva and Mehl.[4] This agreement indicates that the effect of the small amount of porosity developed in copper-alpha brass sandwich-type couples is quite minor.

The values of the intrinsic diffusivities of zinc and copper were then calculated from marker movement and chemical diffusivity data using Eq. (12) and (10) and from marker movement, and area ratios using Eq. (12) and (15). Both methods yielded the same results which are given in Table I. These values refer to the 28 atomic percent zinc composition at the markers.

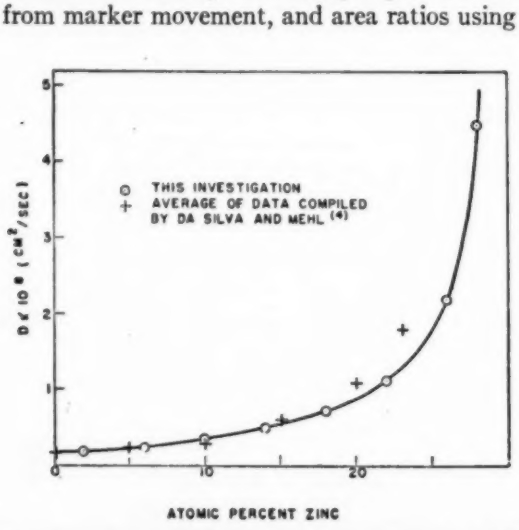


FIG. 5. *D versus* concentration for alpha brass couples at 890°C.

b. Copper-Nickel

Fine alumina particles were placed on the surfaces of $\frac{3}{8}$ in. thick, 1-in. diameter high-purity nickel discs having a grain size of about 1 mm. These specimens were then exposed to copper vapor on both faces for various times in a continuously evacuated furnace at 1060°C. The copper vapor was supplied from similar copper discs which were placed with faces parallel at a spacing of 2 mm. The vapor pressure of copper (2×10^{-4} mm Hg) is about 10^5 that of nickel at this temperature, and the spacing between disks was several orders of magnitude less than the mean free path of the copper atoms in the vacua obtained. Chemical analysis of the copper disks after diffusion proved that negligible nickel was lost from the nickel to the copper.

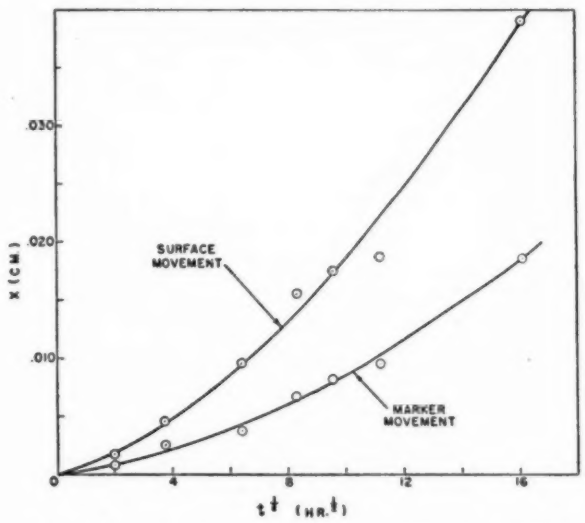


FIG. 6. Surface movement and marker movement *versus* $t^{\frac{1}{2}}$ for copper-nickel couples at 1060°C.

For this system it was obvious from the color of the specimen surfaces that the composition at the vapor-solid interface changed over a considerable period of time. Marker movements were obtained by sectioning the couple and taking the difference between the specimen expansion (surface movement) and depth of burying of the inert markers. The metallographic examination also revealed that no porosity formed anywhere in the diffusion zone. The marker movement did not vary parabolically with time as shown in Fig. 6. Hence it appeared that the boundary conditions were not closely enough satisfied for application of the Matano analysis, and values for D could not be determined from the diffusion penetration curves.

Although the vapor-solid couple was not suitable for measuring D in the copper-nickel system, the difference of intrinsic diffusion coefficients ($D_{Cu}-D_{Ni}$) was calcu-

TABLE I. Intrinsic diffusivities in various systems.

System	Temp (°C)	Composition	Diffusivity difference (D_1-D_2) (cm²/sec)	Chemical diffusivity (D) (cm²/sec)	Intrinsic diffusivities (cm²/sec)	Evaporation rate of volatile component (moles/cm²/sec)	$\left(\frac{\text{Evaporation rate}}{D}\right)$
Alpha brass	890	28.0 atomic percent Zn	5.1×10^{-8}	4.8×10^{-8}	$D_{Zn}=6.3\times10^{-8}$ $D_{Cu}=1.2\times10^{-8}$	1×10^{-2}	2×10^{5}
Cu–Ni	1060	83.0 atomic percent Cu	2.0×10^{-10}	5.0×10^{-10} [a]	$D_{Cu}=3.7\times10^{-10}$ $D_{Ni}=1.7\times10^{-10}$	4×10^{-8}	80
Ag–Au	940	63.5 atomic percent Ag	5.6×10^{-9}	3.8×10^{-9}	$D_{Ag}=7.3\times10^{-9}$ $D_{Au}=1.7\times10^{-9}$	3×10^{-7}	80

[a] See reference 8.

lated from marker movement velocities obtained from Fig. 6, and the penetration curves of the couples diffused for 114 and 259 hours. The composition at the markers in both couples was 82 atomic percent copper. Using the marker velocity data at 114 and 259 hours obtained from Fig. 6 and values of the gradient obtained

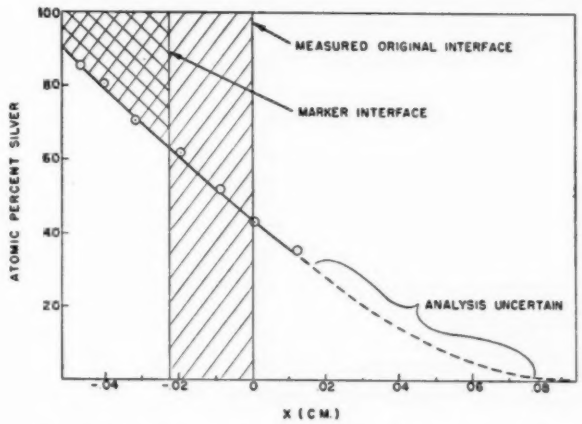


FIG. 7. Diffusion penetration curve of silver-gold couple diffused 64 hr at 940°C.

from the diffusion-penetration curves, $(D_{Cu}-D_{Ni})$ at 83 percent copper was computed. The value of D at this composition was taken from the recent data of Thomas and Birchenall,[8] and the intrinsic diffusivities were calculated using Eqs. (10) and (12). All values are given in Table I.

c. Silver-Gold

High-purity gold disks $\frac{1}{4}$ in. thick and $\frac{3}{4}$ in. in diam. were prepared with a coarse grain size of about 1 mm. Fine tungsten particles were placed on the surfaces of these specimens which were then exposed to silver vapor for various times in a continuously evacuated furnace at 940°C. The silver was supplied from pure silver disks which were placed in an arrangement similar to that already described for the copper-nickel system. The vapor pressure of silver (2×10^{-3} mm Hg) is about 5×10^{3} that of gold at this temperature, and the spacing between disks was several orders of magnitude less than the mean free path of the silver atoms in the vacua obtained. Chemical analysis of the silver disks after diffusion proved that negligible gold was transferred in the opposite direction.

The diffusion penetration curve at 64 hr is shown in Fig. 7. The chemical analysis of slices containing less than 30 atomic percent silver was considered uncertain due to difficulties in separating silver and gold, and, therefore, this portion of the curve is shown dotted. The 90 percent silver surface concentration is somewhat below the desired value of close to 100 percent silver. After 4 hr the surface concentration was found to be 80 percent silver and, apparently, the surface concentration slowly increased during the major portion of the 64-hr period. Hence, the solution $N=f(x/t^{\frac{1}{2}})$ cannot apply strictly, but the curve of marker movement *versus* square root of time, Fig. 8, nevertheless appears linear within experimental error. This observation indicates that an observed apparent parabolic marker movement cannot be used alone as a sensitive criterion that the desired boundary conditions are strictly maintained.

Metallographic examination of the diffusion zones also disclosed the progressive formation of porosity in the region between the specimen surface and the inert

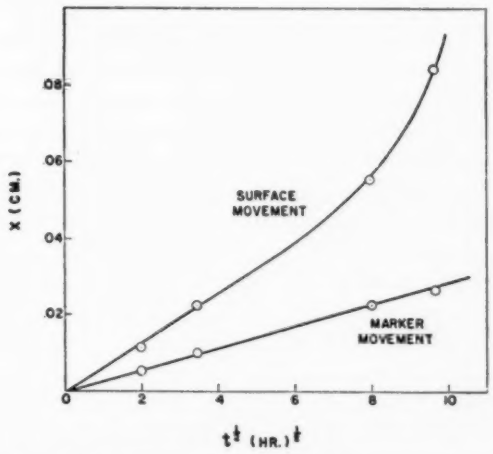


FIG. 8. Surface movement and marker movement *versus* $t^{\frac{1}{2}}$ for silver-gold couples at 940°C.

[8] D. E. Thomas and C. E. Birchenall, J. Metals 4, 867 (1952).

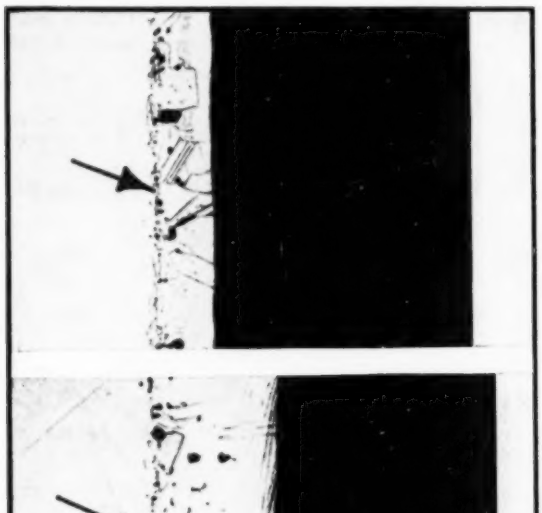


FIG. 9(a). Diffusion zones in vapor-solid silver/gold couples at 940°C. 200×. Above: 4 hr. Middle: 12 hr. Below: 64 hr. (Arrow indicates position of Kirkendall interface.)

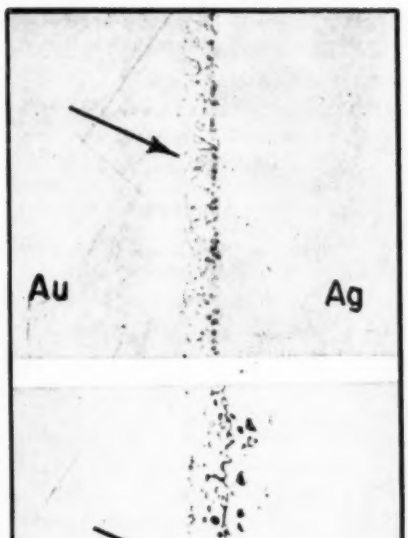


FIG. 9(b). Diffusion zones in sandwich type silver/gold couples at 940°C. 200×. Above: ~1 hr. Middle: 2 hr. Below: 10 hr. (Arrow indicates position of Kirkendall interface.)

markers which produced a deviation from parabolic surface movement at longer times. Since such porosity was found only in the silver-gold vapor-solid couples, a detailed study of its formation was carried out in both vapor-solid and sandwich couples. The sandwich couples containing inert tungsten particles at the original interface were made by hot welding together disks of polycrystalline silver and gold. Typical microstructures are shown in Figs. 9(a) and 9(b).

It is evident that porosity tends to form on the silver-rich side of the Kirkendall interface in both types of couple, and that the amount increases with time. After 64 hr the total volume porosity between the markers, and the surface of the vapor-solid couples was found to be 8 percent by lineal analysis, and after 93 hr 13 percent was found. It is also apparent that the finely distributed porosity formed early in the diffusion cycle next to the Kirkendall interface, is continuously eliminated. This latter observation agrees directly with the conclusions of Seith and Kottmann[9] who have made similar observations in silver-gold sandwich couples.

Even though the boundary conditions were not exactly obeyed, the Matano analysis was applied to the curve of Fig. 7. Porosity effects were still small in the 64-hr specimen and minor corrections§ were made in Fig. 7 to account for the small amount of porosity between the surface and the markers at this time. The resulting D values along with other values obtained by previous investigators, are given in Fig. 10. The agree-

[9] W. Seith and A. Kottmann, Angew. Chem. 64, 379 (1952).

§ For further details see R. W. Balluffi, discussion to reference 8, J. Metals 5, 726 (May, 1953).

ment of our approximate values with previous data is fairly good. It should be remarked here that the results of Ebert and Trommsdorf[10] and Johnson[11] using incremental couples are probably more trustworthy than the values of Seith and Kottmann.[9] The latter values were obtained by the Matano analysis and may be subject to significant porosity errors. Their less dependable values in the silver-rich range have been omitted.

The difference between intrinsic diffusivities ($D_{Ag} - D_{Au}$) at 63.5 atomic percent silver was next calculated from Eq. (12) using the flow velocity and gradient at the marker (see Table I). As previously indicated this method does not depend upon the validity of the solution $N_1=f(\lambda)$ and therefore applies even when variations in the surface composition occur. Using the value of D at the marker composition obtained from Fig. 10, intrinsic diffusivities were calculated by use of Eqs. (10) and (12) and are presented in Table I. Values of D_{Ag} and D_{Au} calculated from Eqs. (12) and (15), using the ratio of diffusivities obtained from the hatched areas in Fig. 7, were the same.

d. Dimensional Changes Normal to the Diffusion Direction

Direct measurements were made of dimensional changes normal to the diffusion direction in the copper-alpha brass and copper-nickel couples. In the alpha-brass system, measurement of the original and final diameters of the couple diffused at 890°C for 35.5 hr proved that normal expansion was less than 0.1 percent. Since the maximum percentage increase of the lattice spacing amounted to about 1.7 percent, this result proves that dimensional changes, even those

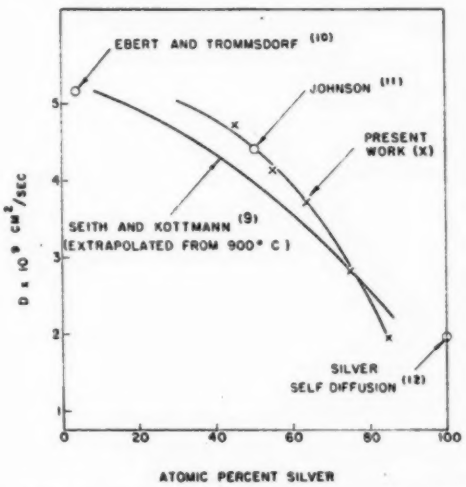


FIG. 10. *D versus* concentration for silver-gold at 940°C. See reference 12 for silver self-diffusion values.

[10] H. Ebert and G. Trommsdorf, Z. Electrochem. **54**, 294 (1950).
[11] W. A. Johnson, Trans. Am. Inst. Mining Met. Engrs. **147**, 331 (1942).
[12] Slifkin, Lazarus, and Tomizuka, J. Appl. Phys. **23**, 1032 (1952).

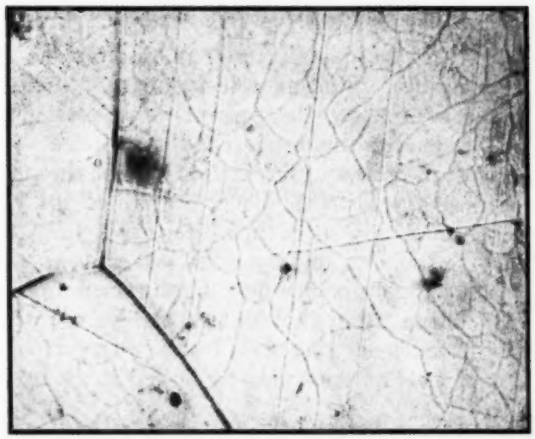


FIG. 11. Polygonization visible on surface of alpha brass couple diffused 35.25 hrs at 890°C. 200×.

resulting from the increase in lattice parameter, are largely restricted to the diffusion direction. This observation agrees with the result of da Silva and Mehl[4] who found no normal dimensional change in the brass side of a massive copper-alpha brass sandwich-type couple.

Similar results were obtained for the copper-nickel couples. A linear expansion of less than 0.2 percent was found in the couple diffused at 1060°C for 259 hr, although the maximum increase of lattice spacing amounted to about 2.5 percent. Normal dimensional changes of the foregoing magnitudes would have negligible effect upon any quantitative calculations and the original assumption of constant specimen cross section seems to be closely approximated.

e. Polygonization in Diffusion Zone

In a previous note,[13] one of the writers has reported metallographic evidence of polygonization in copper-nickel diffusion couples. Upon observation of the surface of a vapor-solid copper-nickel couple directly after diffusion, characteristic grooves at the intersection of polygon boundaries with the surface were plainly evident. Consequently, the final surfaces of the present couples were carefully examined.

Figure 11 shows the surface of the 35.25-hr brass specimen revealing a typical polygonized structure. Several slip lines prove that the orientation difference between adjacent subgrains is quite small. No subgrain structure was visible on the surface of an identical copper disk heated in the absence of zinc vapor.

Careful examination of the surfaces of the silver-gold specimens failed to disclose any visible subgrain boundaries. However, a Laue back reflection x-ray photogram of the surface layer of the specimen diffused 64 hr at 940°C revealed multiple spots, and evidently this specimen was also polygonized.

[13] R. W. Balluffi, J. Appl. Phys. **23**, 1407 (1952).

IV. SUMMARY AND CONCLUSIONS

The present work indicates that absolute measurements of intrinsic diffusion coefficients may be made for binary systems using the present vapor-solid couple method whenever the vapor can be supplied to the specimen surface at a sufficient rate to raise the surface concentration close to the concentration of the vapor source within a very short time. When this condition is satisfied the proper boundary conditions are maintained and chemical diffusivities may be obtained from the diffusion curve using the Matano analysis. Intrinsic diffusion coefficients may then be calculated from the position of inert markers initially placed at the original interface.

This condition is presumably met if the rate of delivery of the volatile metal to the specimens surface is sufficiently high compared to its rate of diffusion into the metal. The ratio of the evaporation rate to the diffusion coefficient may therefore be used as a very approximate empirical criterion for the applicability of this type of couple to various systems. In the present work the technique was successful for the diffusion of zinc into copper at 890°C. At this temperature the ratio is about 2×10^5 (see Table I). The technique was doubtful for the diffusion of silver into gold at 940°C and copper into nickel at 1060°C where the ratios are both about 80.

When the rate of material transport through the vapor phase is too low to maintain a constant surface concentration, vapor-solid couples may still be used to obtain an exact value for the difference of intrinsic diffusivities since this quantity can be calculated from the experimental data without any assumptions concerning the boundary conditions. If values of chemical diffusivity are known from other experiments, the intrinsic diffusivities may then be calculated.

The intrinsic diffusivities in the copper-alpha brass system obtained by the present method are probably more accurate than those derived from sandwich type couples. This is shown quite clearly by the negligible scatter of points representing marker movements in Fig. 4 compared to the broad scatter encountered in the data of da Silva and Mehl.[4] The ratio of D_{Cu} to $D_{Ni}=2.2$ calculated from the present data is considerably greater than the approximate ratio of 1.3 obtained from sandwich type couples in the copper/nickel system by Thomas and Birchenall.[8] It is felt that the use of the vapor-solid couple to determine intrinsic diffusion coefficients insures direct and accurate measurement of marker movements and avoids problems associated with placement of markers and reference foils and with graphical methods of determining the original interface.[4] Porosity formation is also avoided in some systems.

Careful length measurements prove that expansion normal to the diffusion direction is negligible in the present specimens, and it appears that in massive diffusion couples such expansion is restrained by the undiffused bulk of the couple. This conclusion is borne out by the measurements of da Silva and Mehl[4] but is contradicted by Seith and Kottmann.[9] However, there are some discrepancies in the measurements of Seith and Kottmann which are discussed elsewhere.[14] It appears that, after correcting for changes in atomic volume, the assumption of a constant number of atoms per equal increment of distance is closely approximated in massive diffusion specimens.

In agreement with the observations of Barnes[15] the present work indicates that polygonization in the diffusion zone is a frequently occurring phenomenon. Polygonization is probably the result of the mass flow associated with the unequal diffusion processes and may be intensified because the specimen expansion is not isotropic, as it tends to be, but is restricted to the direction of diffusion by the bulk of the couple.

ACKNOWLEDGEMENT

The authors would like to thank Mr. A. Timper and Mr. R. Resnick for performing the experimental work described in this paper.

[14] R. W. Balluffi and L. L. Seigle, Naturwiss. **20**, 524 (1953).
[15] R. S. Barnes, Proc. Phys. Soc. (London) **65B**, 512 (1952).

JOURNAL OF APPLIED PHYSICS VOLUME 25, NUMBER 5 MAY, 1954

Stored Energy and Power Flow in Electron Beams

L. R. WALKER
Bell Telephone Laboratories, Murray Hill, New Jersey
(Received May 11, 1953)

The problem of finding the density of ac energy and the ac power flow in an electron beam, originally uniform in velocity and charge density, which interacts with a weak electromagnetic traveling wave, is considered. For the problem to be significant, the way in which the traveling wave is built up must be specified and the energy and power flow turn out to depend upon the method of establishment.

IN the discussion of traveling-wave tubes and similar electronic devices, one often assumes for simplicity that the electromagnetic fields have the space and time variation of a simple traveling wave. It is the purpose of this note to discuss some considerations which must be taken into account if this picture is used. In small signal theory the amplitude of the fields is supposed to be so small that the equations of motion and continuity may be linearized by the neglect of products of ac quantities. In such a theory the ac velocities and currents in the beam and, hence, the admittance of the latter are readily found. It is of some interest to know the density of stored kinetic energy in such a beam and, also, the flow of ac power. Since these are second-order quantities in the applied ac field amplitude, their evaluation would appear to require the analysis to be carried beyond the usual linear terms. For such terms will contribute to the energy as the product of the second-order density term and the square of the zero-order velocity.

It will be sufficient to consider the simplest problem of this type. Suppose that a beam of electrons, homogeneous in velocity v_0 and charge density ρ_0, interacts with a weak traveling-wave electric field in the direction of its motion (along the positive z axis). Suppose that all forces are longitudinal and of magnitude $-\lambda\cos(\omega t+\Gamma z)$, where λ is small in the sense that it causes small changes in velocity and charge density. The question is now asked, "What is the ac energy density and the ac power flow?" Now it is clear that as thus formulated the question is incomplete. For the applied field extends indefinitely in space and in time, and no connection is established with the "original" homogeneous state of the beam. To make the problem meaningful and also relevant to the kind of application which is usually made of models of this kind, one may suppose that the field has been established in some gradual fashion. Specifically, the force on the electrons will be taken as

$$-\lambda a(z,t)\cos(\omega t+\Gamma z),$$

where $a(z,t)\to 0$ as z or $t\to-\infty$; $a(z,t)\to 1$ as z,t become large and $\partial a/\partial z$, $\partial a/\partial t$ may be made arbitrarily small. Thus the field amplitude varies arbitrarily with position, but builds up very slowly to the amplitude λ for all large z and t. One may now legitimately ask for the average ac energy density and power flow in the beam for large z and t.

Before attempting to answer this question it is of interest to note that the rather meaningless question asked originally may be answered plausibly by using small signal theory uncritically. For, using the usual complex notation of linear theory, one may write

$$\frac{\partial\rho}{\partial t}+\frac{\partial\rho v}{\partial z}=0,$$

$$\frac{\partial v}{\partial t}+v\frac{\partial v}{\partial z}=-\lambda e^{j(\omega t+\Gamma z)},$$

where ρ and v are the charge density and velocity in the beam. λ will be given by eE/m, where e and m are the electron charge and mass and E is the electric field. Writing $\rho=\rho_0+\rho_1 e^{j(\omega t+\Gamma z)}$ and $v=v_0+v_1 e^{j(\omega t+\Gamma z)}$, and making the linear approximations, we have

$$(\omega+\Gamma v_0)\rho_1+\Gamma\rho_0 v_1=0$$

$$(\omega+\Gamma v_0)v_1=j\lambda$$

$$v_1=\frac{j\lambda}{(\omega+\Gamma v_0)}$$

$$\rho_1=\frac{-j\Gamma\rho_0\lambda}{(\omega+\Gamma v_0)^2},$$

and for the rf current, $i_1=\rho_0 v_1+v_0\rho_1$,

$$i_1=\frac{j\omega\rho_0\lambda}{(\omega+\Gamma v_0)^2}.$$

Suppose that the field E does not vary in the x direction and let the beam have a small thickness δ in the y direction. The tangential magnetic fields, H_{x-} and H_{x+}, on the $-y$ and $+y$ sides of the beam satisfy

$$H_{x+}=H_{x-}-\frac{j\omega\rho_0\delta\lambda}{(\omega+\Gamma v_0)^2}$$

$$=H_{x-}-\frac{j\omega\rho_0\delta(e/m)E}{(\omega+\Gamma v_0)^2}.$$

The power entering the beam per unit area is then

$$\tfrac{1}{2}\,\mathrm{Re}[E^*(H_{z^-}-H_{z^+})]=\tfrac{1}{2}\,\mathrm{Re}\left[j\frac{\omega\rho_0\delta(e/m)}{(\omega+\Gamma v_0)^2}\right]|E|^2. \quad (1)$$

The quantity $\omega\rho_0\delta(e/m)/(\omega+\Gamma v_0)^2$ will be called the susceptance of the beam and written as $B(\omega,\Gamma)$. Thus the power absorbed per unit area will be

$$\tfrac{1}{2}\,\mathrm{Re}(jB(\omega,\Gamma))|E|^2$$

and will vanish for real ω and Γ.

Now suppose the ac energy storage per unit length in the beam to be $\frac{1}{2}K|E|^2$. Imagine that ω is slightly complex, so that $\omega=\omega_1+j\omega_2$, while Γ is real. Then

$$B(\omega,\Gamma)=B(\omega_1,\Gamma)+j\omega_2\left(\frac{\partial B(\omega,\Gamma)}{\partial\omega}\right)_{\omega=\omega_1}$$

and

$$\mathrm{Re}jB(\omega,\Gamma)=-\omega_2\left(\frac{\partial B}{\partial\omega}\right)_{\omega=\omega_1}.$$

E however varies as $\exp(j\omega_1-\omega_2)t$ and $|E|^2$ as $\exp-2\omega_2 t$, so that the rate of change of stored energy per unit length is

$$\frac{\partial}{\partial t}\tfrac{1}{2}K|E|^2=-\omega_2K|E|^2,$$

and this must equal the rate at which energy enters the beam. So we have

$$\omega_2K|E|^2=\tfrac{1}{2}\omega_2\left(\frac{\partial B}{\partial\omega}\right)_{\omega=\omega_1}|E|^2$$

and

$$K=\frac{1}{2}\left(\frac{\partial B}{\partial\omega}\right)_{\omega=\omega_1}. \quad (2)$$

This is a result familiar in the ordinary analysis of reactive networks.* By an exactly similar argument keeping ω real and making Γ slightly complex, we find that if $\frac{1}{2}J|E|^2$ is the ac power flow in the beam then

$$J=\frac{1}{2}\left(\frac{\partial B}{\partial\Gamma}\right)_{\Gamma=\Gamma_1}.$$

Returning to the more definitely stated problem it will be convenient to suppose that the undisturbed beam contains, not a single velocity of electrons, but a distribution of velocities. It will be supposed however that all electrons travel in the same direction and that none is synchronous with the velocity $\omega\Gamma$ of the traveling wave. The relative number of electrons in the modulated beam in the element z, $z+dz$ with velocities between v and $v+dv$ is written as $f(z,t,v)dzdv$, so that $\rho_0\int_0^\infty f(z,t,v)dv$ is the charge density at z and t and ρ_0 is the charge density of the undisturbed beam. The function f satisfies

$$\frac{\partial f}{\partial t}+v\frac{\partial f}{\partial z}=\frac{\partial f}{\partial v}\lambda a\cos(\omega t+\Gamma z), \quad (4)$$

for this implies the conservation of the number of electrons in phase space.

Now, put

$$f=f_0(v)+\lambda f_1(z,t,v)+\lambda^2 f_2(z,t,v)+\cdots \quad (5)$$

where $f_0(v)$ is the velocity distribution in the original beam. Substituting in (4) and equating terms of the same order in λ, one finds

$$\frac{\partial f_1}{\partial t}+v\frac{\partial f_1}{\partial z}=\frac{\partial f_0}{\partial v}a(z,t)\cos(\omega t+\Gamma z), \quad (6a)$$

$$\frac{\partial f_2}{\partial t}+v\frac{\partial f_2}{\partial z}=\frac{\partial f_1}{\partial v}a(z,t)\cos(\omega t+\Gamma z), \quad (6b)$$

and so on. To solve these equations f_1 and f_2 are written in the form

$$f_1=A\cos(\omega t+\Gamma z)+B\sin(\omega t+\Gamma z),$$

$$f_2=C+D\cos 2(\omega t+\Gamma z)+E\sin 2(\omega t+\Gamma z),$$

where A, B, C, D, and E are functions of z, t, and v. When these expressions are inserted for f_1 and f_2 and similar trigonometric terms collected the following equations arise: from (6a)

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)A+(\omega+\Gamma v)B=\frac{\partial f_0}{\partial v}a(z,t), \quad (7a)$$

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)B-(\omega+\Gamma v)A=0; \quad (7b)$$

from (6b)

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)C+\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)D\cos 2(\omega t+\Gamma z)$$

$$+\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)E\sin 2(\omega t+\Gamma z)-2(\omega+\Gamma v)$$

$$\times D\sin 2(\omega t+\Gamma z)+2(\omega+\Gamma v)E\cos 2(\omega t+\Gamma z)$$

$$=a(z,t)\cos(\omega t+\Gamma z)\left[\frac{\partial A}{\partial v}\cos(\omega t+\Gamma z)\right.$$

$$\left.+\frac{\partial B}{\partial v}\sin(\omega t+\Gamma z)\right],$$

* An additional factor $\frac{1}{2}$ appears here because of the space average involved.

or

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)C=\tfrac{1}{2}a(z,t)\frac{\partial A}{\partial v} \tag{8a}$$

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)D+2(\omega+\Gamma v)E=\tfrac{1}{2}a(z,t)\frac{\partial A}{\partial v} \tag{8b}$$

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)E-2(\omega+\Gamma v)D=\tfrac{1}{2}a(z,t)\frac{\partial B}{\partial v}. \tag{8c}$$

Eliminating B from Eqs. (7a) and (7b) one obtains

$$\left[\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)^2+(\omega+\Gamma v)^2\right]A=\frac{\partial f_0}{\partial v}\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)a(z,t).$$

The right-hand side of this equation is small because the rate of change of a is small and one may write

$$A=\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)a(z,t)-\frac{1}{(\omega+\Gamma v)^2}\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)^2 A. \tag{9}$$

It is clear that A may be expanded in a series of the form

$$\sum\frac{A_n}{(\omega+\Gamma v)^{2n}}\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)^{2n-1}a(z,t)$$

and this will converge rapidly if $\omega+\Gamma v\neq 0$, since

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)^{2n-1}a(z,t)$$

may be chosen to be as small as one desires. In a similar fashion it is seen that C, D, and E are also very slowly varying functions of z and t. Thus, when average energy densities and power flows are calculated only the term C in f will contribute, all the others being attached to trigonometrically varying terms and thus averaging nearly to zero. Now C has to satisfy

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)C=\frac{a(z,t)}{2}\frac{\partial}{\partial v}\left[\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)a(z,t)\right]+\text{smaller terms}$$

$$=\frac{a(z,t)}{2}\left[\frac{\partial}{\partial v}\left(\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\right)\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)a(z,t)+\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\frac{\partial a}{\partial z}\right]+\cdots. \tag{10}$$

Thus,

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)\left(C-\frac{1}{4}\frac{\partial}{\partial v}\left[\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\right]a^2(z,t)\right)=\frac{\partial f_0/\partial v}{4(\omega+\Gamma v)^2}\frac{\partial a^2}{\partial z}+\cdots. \tag{11}$$

The solution of

$$\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial z}\right)F(z,t,v)=G(z,t,v), \tag{12}$$

which tends to zero as t and z tend to $-\infty$ is given by

$$F(z,t,v)=\int_0^\infty G(z-v\sigma,t-\sigma,v)d\sigma.$$

This may easily be established by changing from the variables, z and t, to $z-vt$ and t in the Eq. (12). Thus, C is given by

$$C=-\frac{1}{4}\frac{\partial}{\partial v}\left[\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\right]a^2(z,t)+\frac{\partial f_0/\partial v}{4(\omega+\Gamma v)^2}\int_0^\infty\frac{\partial a^2}{\partial z}(z-v\sigma,t-\sigma)d\sigma, \tag{13}$$

and for large z and t

$$C=-\frac{1}{4}\frac{\partial}{\partial v}\left[\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\right]+\frac{\partial f_0/\partial v}{4(\omega+\Gamma v)^2}\lim_{z,t\to\infty}\int_0^\infty\frac{\partial a^2}{\partial z}(z-v\sigma,t-\sigma)d\sigma. \tag{14}$$

The fact that this expression for C depends explicitly upon the form of the function $a(z,t)$ indicates that the problem which has been set does not have a solution which is independent of the way the field is built up. In general, then the stored energy and power flow are not uniquely determined by the first-order state of the beam.

Since the stored ac energy density of the beam is proportional to $\frac{1}{2}\int_0^\infty f_2v^2dv$ and the ac power flow to $\frac{1}{2}\int_0^\infty f_2v^3dv$, it is convenient to consider $\frac{1}{2}\int_0^\infty f_2v^ndv$, which will be written as $\langle v^n\rangle$. Suppose that the two extreme cases are now considered; first, that in which $a(z,t)$ has no z variation and, second, that in which it has no t variation. Write $\langle v^n\rangle_1$ and $\langle v^n\rangle_2$ for the integrals in these cases. Then

$$\langle v^n\rangle_1=-\frac{1}{4}\int_0^\infty v^n\frac{d}{dv}\left[\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\right]dv$$

$$=-\frac{1}{4}\int_0^\infty dv\frac{d}{dv}\left[\frac{v^n df_0/dv}{(\omega+\Gamma v)^2}\right]-\frac{n}{4}\int_0^\infty\frac{v^{n-1}df_0/dv}{(\omega+\Gamma v)^2}dv$$

$$=-\frac{n}{4}\int_0^\infty\frac{d}{dv}\left(\frac{v^{n-1}f_0}{(\omega+\Gamma v)^2}\right)dv+\frac{n}{4}\int_0^\infty f_0(v)\frac{d}{dv}\left(\frac{v^{n-1}}{(\omega+\Gamma v)^2}\right)dv$$

$$=\frac{n}{4}\int_0^\infty f_0(v)\frac{d}{dv}\left(\frac{v^{n-1}}{(\omega+\Gamma v)^2}\right)dv, \tag{15}$$

provided $f_0(v)$ tends to zero sufficiently rapidly at 0 and ∞. In the second case, since

$$\frac{\partial a^2}{\partial z}(z-v\sigma,t-\sigma)=\frac{\partial a^2}{\partial z}(z-v\sigma)=-\frac{1}{v}\frac{\partial a^2}{\partial \sigma}(z-v\sigma),$$

one has

$$\lim_{z,t\to\infty}\int_0^\infty \frac{\partial a^2}{\partial z}(z-v\sigma,t-\sigma)d\sigma=-\left[\frac{1}{v}a^2(z-v\sigma)\right]_0^\infty$$

$$=-\frac{1}{v}\lim_{z\to\infty}a^2(z)=\frac{1}{v}$$

and

$$\langle v^n\rangle_2=\frac{1}{4}\int_0^\infty v^n\frac{d}{dv}\left[\frac{\partial f_0/\partial v}{(\omega+\Gamma v)^2}\right]dv+\frac{1}{4}\int_0^\infty v^{n-1}\frac{df_0/dv}{(\omega+\Gamma v)^2}dv$$

$$=\frac{n-1}{4}\int_0^\infty f_0(v)\frac{d}{dv}\left[\frac{v^{n-1}}{(\omega+\Gamma v)^2}\right]dv. \tag{16}$$

Thus, generally,

$$\langle v^n\rangle_2/\langle v^n\rangle_1=(n-1)/n.$$

What is the relation of these results to those obtained from the linear theory? The expression for the stored energy was derived on the assumption that the field had a fixed space periodicity and should therefore correspond to $\langle v^2\rangle_1$. Now,

$$\langle v^2\rangle_1=\frac{1}{2}\int_0^\infty f_0(v)\frac{d}{dv}\left(\frac{v}{(\omega+\Gamma v)^2}\right)dv$$

$$=\frac{1}{2}\int_0^\infty f_0(v)\left[\frac{1}{(\omega+\Gamma v)^2}-\frac{2\Gamma v}{(\omega+\Gamma v)^3}\right]dv$$

$$=\frac{1}{2}\int_0^\infty f_0(v)\frac{d}{d\omega}\left[\frac{\Gamma v}{(\omega+\Gamma v)^2}-\frac{1}{(\omega+\Gamma v)}\right]dv$$

$$=+\frac{1}{2}\int_0^\infty f_0(v)\frac{d}{d\omega}\left(\frac{-\omega}{(\omega+\Gamma v)^2}\right)dv. \tag{17}$$

This agrees with the earlier result since the total stored energy of the distribution is the sum of the stored energies of the separate streams. The signs and numerical factors are correct by the following considerations. The energy density is $\lambda^2 m\langle v^2\rangle$ multiplied by the number of electrons per cc. λ is e/mE and the number density is $-\rho_0/e$. Thus the energy density is $-(e\rho_0/m)\langle v^2\rangle E^2$ and the beam has a volume δ per unit area, giving an energy storage.

$$\frac{E^2}{2}\int_0^\infty f_0(v)\frac{d}{d\omega}\left[\frac{e\rho_0\omega/m\delta}{(\omega+\Gamma v)^2}\right]dv.$$

Similarly the earlier expression for the power flow was based upon the field's having a constant time periodicity and might coincide with $\langle v^3\rangle_2$. But,

$$\langle v^3\rangle_2=\frac{1}{2}\int_0^\infty f_0(v)\frac{d}{dv}\left(\frac{v^2}{(\omega+\Gamma v)^2}\right)dv$$

$$=\frac{1}{2}\int_0^\infty f_0(v)\left(\frac{2v}{(\omega+\Gamma v)^2}-\frac{2\Gamma v^2}{(\omega+\Gamma v)^3}\right)dv$$

$$=\frac{1}{2}\int_0^\infty f_0(v)\left(\frac{2\omega v}{(\omega+\Gamma v)^3}\right)dv$$

$$=+\frac{1}{2}\int_0^\infty f_0(v)\frac{\partial}{\partial\Gamma}\left(\frac{-\omega}{(\omega+\Gamma v)^2}\right)dv. \tag{18}$$

This is again in agreement with the earlier result.

The conclusions to be drawn from this analysis, then, are that for a traveling wave field established in a variety of ways the stored kinetic energy and power flow will be different from one case to another. Further, linear theory will give correct results if correctly applied.

It should be noted that the expressions for $\langle v^2\rangle_1$ and $\langle v^3\rangle_2$ indicate that the stored energy or power flow may be positive or negative. Negative stored energy implies that work was extracted from the original dc beam in bringing it to its modulated condition. Clearly a beam with negative stored energy is potentially unstable in the sense that it could pass to a state of greater modulation by doing work on its surroundings. $\langle v^2\rangle_1$ for a beam of velocity v is proportional to $(\omega-\Gamma v)/(\omega+\Gamma v)^3$ and if $-\omega/\Gamma$ is written as v_w, the wave velocity, it follows that

$$\langle v^2\rangle_1\propto\frac{v_w^2}{\omega^2}\cdot\frac{v+v_w}{(v_w-v)^3}.$$

$\langle v^2\rangle_1$ is negative then, for positive v_w, if $v<-v_w$ or $v>v_w$. The second case is the more interesting since the negative stored energy can be larger for this case. The energy is negative here if the wave velocity is less than the electron velocity and positive if the wave velocity is the greater.

JOURNAL OF APPLIED PHYSICS VOLUME 25, NUMBER 5 MAY, 1954

Information Theory and Knowledge

D. K. C. MacDonald
Division of Physics, National Research Council, Ottawa, Canada
(Received September 21, 1953)

Distinction between data transmission from one person (or point) to another and the inception of new knowledge is analyzed in relation to information theory. The customary measure $H=-\Sigma_i p_i \log p_i$ applies essentially to the former case, and after discussing the requirements appropriate to a yardstick for new knowledge an elementary measure for this concept is tentatively proposed. The intuitive requirements, however, are not all mutually compatible and the present measure can only be regarded as provisional; in particular some "memory" characteristics appear mandatory for future development.

INTRODUCTION

In a provocative paper recently, Brillouin[1] is led to suggest that we may have to consider two different *kinds* of information. The first he calls: *Absolute Information* "which exists as soon as one person has it, and should be counted as the same given amount of information, whether it is known to one man or to millions." The second is termed: *Distributed Information* which he defines as ". . . the product of the amount of absolute information and the number of people who share that information," and he goes on to say ". . . Distribution of information . . . increases the amount of distributed information without changing the total absolute information available."

Expressed otherwise, it appears to me that Brillouin wishes to distinguish, on the one hand, between new knowledge and, on the other, the transfer of data from one person (or point) to another. The difficulty of recognizing and resolving this distinction has indeed made itself felt at various times before now, but I believe that the correct solution has not yet been found; this paper makes a suggestion towards that end. Information theory, as brought to the fore by Shannon[2] particularly, presents to us the quantity

$$H=-\Sigma_i p_i \log_2 p_i \qquad (1)$$

as a measure of the amount of information in units of "bits" obtainable from some source, and this seems admirably suited (being indeed designed for the purpose) for the description of the transfer of data within some prescribed frame of reference from one point to another. It appears however that the measure H is *not* suited for the assessment and discussion of new *knowledge*, and we shall try here to find a more appropriate measure for this latter "quantity." Apart from any academic or aesthetic considerations, this can only be of ultimate value if we find that such a measure can be used numerically to assess satisfactorily our gain or otherwise in some change of situation of primary knowledge; and this, at heart, must agree qualitatively with our elementary intuition. We have seen with great interest Gabor and Gabor's (1953) essay: "A Mathematical Theory of Freedom," wherein a numerical measure has been proposed for the effective freedom of action of a population based on *a posteriori* observation of the range of choice exerted by the population. While, as the authors themselves emphasize, so sophisticated a concept requires great care in any numerical interpretation, its possible quantitative application to social and economic problems such as effective electoral liberty, for example, appears most attractive.

THE DISTINCTION BETWEEN DATA ("INFORMATION") AND "NEW" KNOWLEDGE

Let us consider a situation used to illustrate the conventional field of information theory. We suppose that we have a power station together with its generators, etc., situated in some isolated place; it is then very familiar that we can transmit electrical signals to the substation which will start, stop, or reverse the generator, control its speed, and so on. This may then of course be compared with the process of sending verbal orders, or messages, from some transmitting station to a receiving post. In addition, however, to complete the communication link we should like the generator to be able to "talk back" to us and this may quite readily be achieved in the sense that when the machine has in fact responded to a "request" from our end to start up, a signal can be sent back automatically as soon as, say, the armature has attained a prearranged speed . . . and so on. The communication link can readily be broadened further to permit us to be informed if, say, faults develop in the generating system. Thus, for example, a lightning surge on one of the circuits would cause an over-voltage relay to "trip" temporarily, and a distinctive signal would then be sent to us, while, if no serious damage had occurred, the relay, on resetting automatically, would initiate a second appropriate signal. This framework is obviously susceptible to wide ramification and we are certainly fully justified in saying that we are "in communication" with the substation and that data (or information in the Shannon sense of the word) is being transmitted to and fro. All the usual calculations may be made (and applied) of channel capacity, optimum coding, etc.

[1] L. Brillouin, Am. Scientist **38**, 594 (1950).
[2] C. E. Shannon, Bell System Tech. J. **27**, 376, 623 (1948); **30**, 50 (1951).

However, it appears evident to the writer that so long as this communication system *is* working normally and fulfilling its expected function, no essentially new *knowledge* is being created. Thus, in order to design an appropriate communication system and make it function efficiently, we must have first a knowledge (the word now being used primarily in its intuitive sense!) of the characteristic functions of the generator, etc., and more particularly of the faults which may occur therein.

Given these and their expected relative frequencies of occurrence, say p_i for the ith "fault," etc., we can then proceed to design our system and use the familiar measure of information content per signal sent:

$$H = -\Sigma_i p_i \log_2 p_i (\text{bits}) \qquad (2)$$

to analyze its performance quantitatively.

If now in any "system"—ranging from the signalling of messages by a broken-down motor at one end of the scale to the delivery of lectures on, say, quantum field theory, at the other—the statistics of the "transmitter" (motor, lecturer) in the sense of the p_i's for all possible messages are truly known, then there is really no possibility of creating any new knowledge in the intuitively fundamental sense. All that one can do is to give (send) notification of the existence of situations already fully visualized and whose statistics are already known. Of course the probable trend of events will even be known beforehand if the communication system is to be used at high efficiency since then we shall have to know not merely the individual probabilities of the enumerable situations which may occur but also all the dependent probabilities: $p_{i;j}$, $p_{ij;k}$, etc. (which are necessary for appropriate coding).

This last conclusion suggests already a possible relationship of new knowledge within the established framework as opposed to "mere" data transmission with the *fluctuations* of a data-transmission system from its most probable (average) behavior. We shall go further here, however, and suggest that it is only when the already-envisaged or established system is actually *unable to cope* with a situation which presents itself that any essentially *new* knowledge (or "absolute information" to use Brillouin's terminology) is generated. Thus, to revert to our previous examples, a *new* fault, not previously encountered, occurs in our generator for which therefore *no* signal has been provided, or our lecturer has to attempt to introduce and define new words or to ascribe new meanings to words in fresh contexts not previously envisaged by the audience.

A number of illustrative examples come readily to mind in this connection. It appears true to say that the science of mathematics only really progresses when we are forced to invent new symbols and/or to increase our vocabulary of functions and that significant progress in physical science only occurs when we can no longer adequately describe the behavior of observed systems within the existing terminology. If indeed the existing description proves entirely adequate in some field of investigation, we can go so far as to say that *further* research in that field is surely quite sterile. It is perhaps unwise to venture further afield, but we can at least notice that in the artistic domains, such as literature and music, the same trend is obvious when it becomes felt that existing modes of presentation are no longer adequate.

A MEASURE FOR NEW KNOWLEDGE

One can summarize the over-all process of evolution of natural science and the increase of *knowledge*, in terms of a "communication system" (see Fig. 1) as follows.

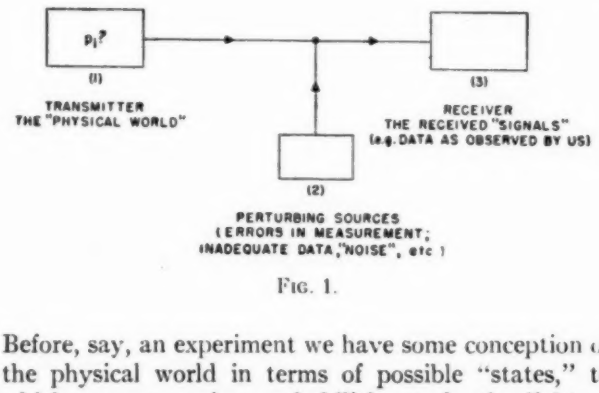


FIG. 1.

Before, say, an experiment we have some conception of the physical world in terms of possible "states," to which we may assign probabilities p_i in the light of our knowledge to date. After the experiment, and/or the analytical examination of available data, we shall in general make some reassessment of the p_i's in the light of our increased knowledge. [Such an interpretation has indeed been applied quantitatively before to a specific experimental problem such as the analysis of received radar data, inevitably distorted to some degree by random noise (Woodward;[3] Woodward and Davies).[4]] In this sense we may of course say that the problem of physical science is essentially one in *a posteriori* probabilities (cf. e.g., Fisher,[5] 1935).

We therefore ask: if the probabilities *before* an "experiment" (in its most general sense) are p_i and *afterwards* are P_i how should we assess numerically our gain in knowledge on some suitable scale of units? Let us emphasize immediately that: first, it is *not* our province here to state *how* the p_i's and P_i's are to be assessed,* any more than it is the province of conventional information theory to *assess*, in general, the p_i's of the signals in a transmitting source; and secondly, that there will be no *unique* measure of this kind—all that we can seek to provide (as with Shannon's H) is an

[3] P. M. Woodward, Proc. Inst. Radio Engrs. 39, 1521 (1951).
[4] P. M. Woodward and I. L. Davies, Proc. Inst. Elec. Engrs. (Pt. III) 99, 37 (1952).
[5] R. A. Fisher, *The Design of Experiments* (Oliver and Boyd, Edinburgh, 1935).
* For this we might turn again, for example, to Fisher (reference 5), Chap. XI in particular.

objective numerical measure on a suitable scale which will agree as far as possible with our intuition and then, we hope, be useful for further discussions. Finally, our measure can only be *relative* to the situation in hand; thus, it would be meaningless to ask whether the statement: "the atomic weight of potassium is 39.1": represents more (or less) knowledge than: "there is approximately one free electron per atom in sodium metal" just as it is meaningless in general to ask whether a sonata by Beethoven is "better" (or "worse") than a picture by Picasso.

In order, then, to formulate a suitable expression we must first suggest the requirements considered desirable that ΔI—the change in "new information" or "knowledge"—should fulfil.

(1) On the assumption that our "experiments" are always "sincere" (i.e. that no *deliberately* false data is inserted), ΔI should never be negative.

(2) ΔI should be monotonically ordered: that is to say, a change of probabilities: $(\frac{1}{4}, \frac{3}{4})$ to $(\frac{3}{4}, \frac{1}{4})$, for example, should yield a greater ΔI than $(\frac{1}{4}, \frac{3}{4})$ to $(\frac{1}{2}, \frac{1}{2})$.

(3) ΔI should be symmetrical: that is, a change from $(\frac{1}{3}, \frac{2}{3})$ to $(\frac{2}{3}, \frac{1}{3})$ should yield the same ΔI as the inverse change.

(4) ΔI should be linearly additive in the sense that the successive changes $(\frac{1}{4}; \frac{3}{4}) \rightarrow (\frac{1}{2}, \frac{1}{2}) \rightarrow (\frac{3}{4}, \frac{1}{4})$ should yield the same result as the single change $(\frac{1}{4}; \frac{3}{4})$ to $(\frac{3}{4}; \frac{1}{4})$.

(5) A logarithmic basis appears desirable.

(6) For analytical convenience, we should like to exclude the modulus sign as an artifice to ensure condition (1), and the expression should be as "simple" in form as possible.

It is readily seen that these conditions are not all mutually compatible. Thus by conditions (3) and (1) the null change $(p_i;p_j) \rightarrow (p_i;p_j)$ must yield $\Delta I=0$, while by (1) the successive changes: $(p_i;p_j) \rightarrow (p_i';p_j') \rightarrow (p_i;p_j)$ (where $p_i' \neq p_i$) must yield an over-all $\Delta I>0$, since each change must give a *positive* ΔI, and this then evidently contradicts condition (4). A possible measure satisfying all conditions except (4), is then:

$$\Delta I=\sum_i (\Pi_i-\pi_i) \log\left(\frac{\Pi_i}{\pi_i}\right) \tag{3}$$

where $\Pi_i=1-P_i$; $\pi_i=1-p_i$ and the π's may be termed the "improbabilities." The choice of this form rather than the more immediately obvious

$$\Delta I=\sum_i (P_i-p_i) \log\frac{P_i}{p_i} \tag{4}$$

will be justified below. The units will be "bits" if the logarithms are to base 2, or "nits"[6] if natural logarithms are employed. The limitation of our measure for ΔI in (3), in contradicting condition (4), is illustrated by comparing on the one hand, *successive* transitions, say $(\pi_{i1}, \cdots) \rightarrow (\pi_{i2}, \cdots) \rightarrow (\pi_{i3}, \cdots)$, with the *direct* change $(\pi_{i1}, \cdots) \rightarrow (\pi_{i3}, \cdots)$. In these cases the *difference* of ΔI is given by:

$$\delta \equiv \Delta I_{\text{direct}} - \Delta I_{\text{successive}}$$

$$=\sum_i \left\{ (\pi_{3i}-\pi_{2i}) \log\left(\frac{\pi_{2i}}{\pi_{1i}}\right) + (\pi_{2i}-\pi_{1i}) \log\frac{\pi_{3i}}{\pi_{2i}} \right\}. \tag{5}$$

(It is then obvious that $\delta=0$ if either $(\pi_{2i}, \cdots) = (\pi_{1i}, \cdots)$ or $(\pi_{3i}, \cdots) = (\pi_{2i}, \cdots)$, reducing the successive transition to a direct one.) Restricting ourselves, for simplicity, to the simplest case of dual probabilities (π_a, π_b), say, it then follows readily that if the transitions are monotonic, e.g., $\pi_{3a}>\pi_{2a}>\pi_{1a}$, then $\delta>0$. This would then suggest that more new knowledge is gained in this case by a single "experiment" yielding the transition, $\pi_{1a} \rightarrow \pi_{3a}$, than by two successive "experiments." Conversely, a nonmonotonic change (e.g., $\pi_{3a}<\pi_{2a}>\pi_{1a}$) yields $\delta<0$. These implications do not, however, seem immediately acceptable to the intuition and indeed may well prove untenable.

Some consideration of this situation suggests that a measure which did not merely take account of a *single* change of probabilities, but rather included some "memory" of previous transitions would be more nearly ideal. In more mathematical terms it appears desirable to recognize "increase of knowledge" as a *non*-Markoffian process. It seems interesting that we are analytically driven to this conclusion since it would certainly appear *intuitively* agreeable that our increase of knowledge would depend not merely on the *immediately previous* assessment of the situation but on how this was earlier approached.

It is evident from (3) that if any π_i (say π_k) becomes precisely zero (and hence all other π_i's become unity), ΔI becomes infinite. This corresponds to the achievement of *certainty* (a "singular point") in our assessment of the probabilities ($p_k=1$, all other $p_i=0$) and it is for this reason that (3) is preferred to (4), as remarked earlier. At first it appeared somewhat surprising that a transition, say, (9/10, 1/10)→(1,0) should yield $\Delta I=\infty$, but indeed this became intuitively very satisfactory because physically such a transition will correspond to the recognition of some *natural law*. The achievement of a natural law implies a certainty about the outcome of all *further* experiments in this field and consequently the infinite assessment is entirely agreeable. It is also worth noting, however, that the logarithmic dependence involves a very *slow* approach to infinity so that even (999/1000, 1/1000) is still "a long way" from (1, 0), underlining the importance of the vital step to the certainty of the physical *law*. This evidently concords with Brillouin's appreciation of the

[6] D. K. C. MacDonald, J. Appl. Phys. **23**, 529 (1952).

apparent informational dilemma involved in the discovery and publication of a natural law. We may now indeed suggest that in this field lies an essential difference between the human brain and any "calculating machine" or "electronic brain." A machine, given an accumulation of experimental data, etc., would be able to make certain quite precise *deductions* about the probability of further experiments, etc., but only the human brain can make the synthetic and *inductive generalization* involved in the statement of a natural law. Of course, if one does not wish to admit that there *is* any essential difference between the human brain and the machine, one may point to the fact that natural laws, as they have been enunciated in the past, (e.g. Newton's laws) do not appear to stand the test of time indefinitely! . . . The debate continues.

CONCLUSION

We may perhaps sum up by saying that any communication system (in its most general sense) which has reached a state of "perfection" where all "redundancy" (or lack of efficiency as a conventional communication system) has been eliminated is impotent to produce new knowledge. While necessity may be the mother of invention it appears that efficiency is not always the father of knowledge.

Limiting Pressure on Hydrofoils at Small Submergence Depths

E. V. LAITONE*
University of California, Berkeley, California
(Received August 31, 1953)

It is shown that the magnitude of the local velocity and the surface pressure coefficient can be limited over the entire upper surface of a two-dimensional hydrofoil as it approaches the free water surface. In this limiting condition the flow over the upper surface is governed by the shallow water wave propagation phenomenon, and the limiting pressure or velocity is determined by the maximum depth attainable from the hydraulic jump relations. The experimental data verifies this new theory and proves that the previous two-dimensional hydrofoil theories are valid only for relatively deeply submerged bodies.

INTRODUCTION

THE experimental data shown in Fig. 1 was obtained by Ausman,[1] who pointed out the apparent similarity of the limiting pressure coefficient (found on the upper surface of the hydrofoil when near the free surface) to the limiting value found when cavitation occurs. However, the magnitudes of the cut-off or limiting pressure coefficients shown in Fig. 1 are much less than those necessary for cavitation (when the local static pressure must drop to its absolute minimum equal to the vapor pressure of the water[2]).

It will now be shown that this limiting pressure over a two-dimensional hydrofoil at small submergence depths is directly related to the phenomenon of shallow water wave propagation. This is physically evident if one considers a thin hydrofoil near the surface. Since its chord is relatively long as compared to its water depth, therefore, the flow over its upper rear surface must be similar to the shallow water flow in an inclined channel. This analysis is very important since it verifies the experimental data presented which shows that the lift of a large chord two-dimensional hydrofoil becomes negligible as it approaches the free water surface. This is contrary to the prediction of all previous theories on the action of a hydrofoil[3] which are, therefore, in error for a finite chord hydrofoil as its submergence depth becomes small.

This free water surface flow resistriction is illustrated in Fig. 2 where the experimental data show how the maximum negative values of the pressure coefficient decrease with the submergence depth once the theoretical limiting value (solid line) has been reached. The experimental values presented in Fig. 2 are the maximum negative values of the pressure coefficient

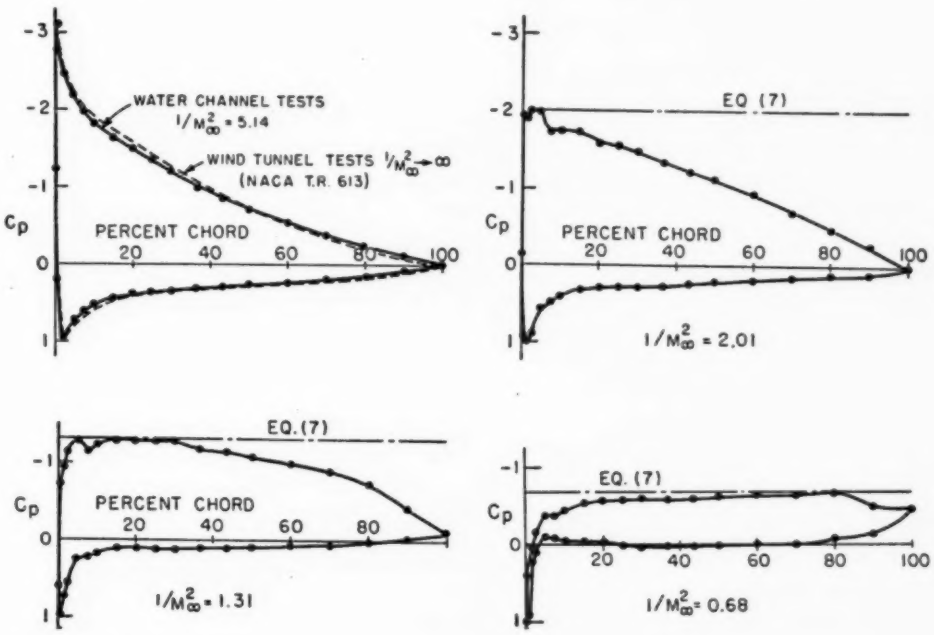


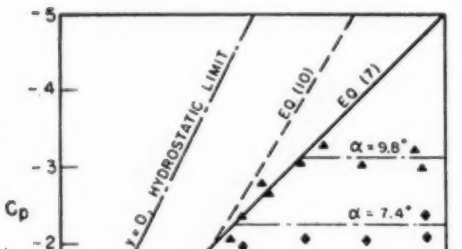
FIG. 1. Two dimensional pressure distribution on NACA 4412 profile at various submergence depths at constant forward speed, $V_\infty = 2.93$ ft/sec, Reynolds number $= 1.4 \times 10^6$, $\alpha = 10°$, $c = \frac{1}{2}$ ft, $h = 1.86$ ft.

* Associate Professor, Mechanical Engineering.

[1] J. S. Ausman, "Experimental Investigation of the Influence of Submergence Depth Upon the Wave-Making Resistance of a Hydrofoil." Master of Science thesis, University of California, Berkeley, (1952).

[2] H. Rouse, *Fluid Mechanics for Hydraulic Engineers* (McGraw-Hill Book Company, Inc., New York, 1938), pp. 388–390, 129, 369.

[3] M. V. Keldysch, and M. A. Lavrentiev, "On the Motion of an Aerofoil Under the Surface of a Heavy Fluid." Tech. Notes of the A.G.I. (Central Aerodynamics Institute), U.S.S.R.



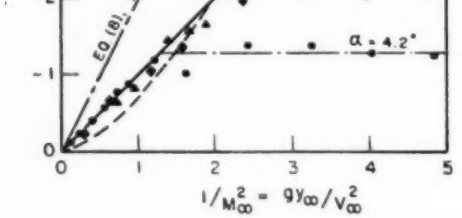


FIG. 2. Comparison of limiting pressure coefficient theory with experimental data at varying water flow and varying depth. Horizontal asymptotes are from wind tunnel tests (Natl. Advisory Comm. Aeronaut. Tech. Rept. 613) at average Reynolds number. Definition of y_* and a typical surface profile are shown below. $V_\infty < \sqrt{gh}$.

found on the upper surface of a hydrofoil at various submergence depths and angles of attack. The limiting values always occurred over a major portion of the upper surface, in the same manner as depicted in Fig. 1.

SYMBOLS

α = angle of attack of hydrofoil, measured from chord line joining trailing edge to furthermost point of leading edge.

$\rho = 62.4/32.2$ lb sec^2/ft^2, mass density of water.

$g = 32.2$ ft/sec^2, gravitational constant.

$M = \dfrac{V}{\sqrt{gy}}$, local Froude number.

V = ft/sec, local velocity.

y = ft, local water depth.

p = lb/ft^2, local static pressure.

$C_p = \dfrac{p - p_\infty}{\frac{1}{2}\rho V_\infty^2}$, local pressure coefficient.

c = ft, hydrofoil chord line length

h = ft, depth of channel containing hydrofoil.

SUBSCRIPTS

0 = stagnation conditions from Bernoulli equation.

1 = limiting value, or value ahead of hydraulic jump.

2 = value after hydraulic jump.

∞ = free stream conditions, y_∞ being defined as depth of trailing edge below free surface.

$*$ = critical condition where $V_* = \sqrt{gy_*}$.

THEORY

If we have a large chord thin hydrofoil at small angles of attack, then as the hydrofoil approaches the free surface of infinitely deep water, we are justified in assuming that all the flow at and aft of the critical depth over the upper surface follows the hydrostatic relations; that is, the centrifugal or vertical accelerations are negligible so the velocity at each section is uniform over the entire water depth. The critical depth is defined the same as in open channel flow;[2] namely,

$$y_* = \frac{V_*^2}{g} \quad \text{or} \quad V_* = \sqrt{gy_*}, \tag{1}$$

so that the Froude number defined by

$$M^2 = \frac{V^2}{gy} \tag{2}$$

is unity at the critical depth (y_*).

Immediately aft of the critical depth, the flow will always be supercritical or rapid ($M > 1$) as long as the hydrofoil is at a positive angle of attack; consequently, the only means of recovering the pressure or water depth is by means of a hydraulic jump. The hydraulic jump[2] reduces the velocity and increases the depth ($y_1 < y_*$ to $y_2 > y_*$) by converting the flow from supercritical to subcritical with some loss of energy. The ratio of the water depths is given by[2]

$$\frac{y_2}{y_1} = \tfrac{1}{2}\left[(1 + 8M_1^2)^{\frac{1}{2}} - 1\right] \tag{3}$$

for $M_1 > 1$.

It is seen that $y_2 = y_1$ for critical flow ($M_1 = 1$) and that y_2/y_1 increases without limit as M_1 increases. However, the absolute value of y_2 is limited to a maximum of 0.8 y_0, where y_0 is the upstream total head (equivalent to the isentropic stagnation conditions ahead of the jump) defined by the Bernoulli equation for hydrostatic flow as,

$$\frac{y_0}{y_1} = \frac{y_1 + \dfrac{V_1^2}{2g}}{y_1} = 1 + \frac{M_1^2}{2}. \tag{4}$$

Then if we solve the equation

$$\frac{d(y_2/y_0)}{d(M_1^2)} = 0,$$

we find the maximum value of y_2/y_0 occurs when

$$M_1^2 = 3, \quad \text{giving} \quad (y_2/y_0)_{max} = 0.8.$$

Corresponding to these conditions we have (using the relations given in reference 2)

$$y_1/y_0=0.4, \quad y_2/y_1=2=V_1/V_2,$$
$$M_2^2=\tfrac{3}{8}, \quad \Delta y_0/y_0=-0.05.$$

The last expression shows that the loss of total head is only 5 percent, so the flow is practically a Bernoulli flow without any losses.

Bakhmeteff[4] has shown that for $M_1=\sqrt{3}$ there is no abrupt jump, and when the actual rise in inches is small, the whole surface wherein the so-called hydraulic jump takes place may be smoothly continuous for a length slightly longer than 5 times the actual rise. The jump is actually unstable for $M_1<\sqrt{3}$, and forms a very abrupt discontinuity for $M_1>\sqrt{3}$ (see p. 249, reference 4 or p. 389, reference 2). When $M_1>\sqrt{3}$ the energy loss increases rapidly with M_1 and becomes extremely large (see p. 390, reference 2). Bakhmeteff[4] also emphasizes that experimentally the small slopes cannot produce the larger hydraulic jumps corresponding to $M_1>\sqrt{3}$.

Therefore, if the flow over the upper surface of a hydrofoil has become hydrostatic as it approaches the free water surface, it must be limited by the following conditions:

$$M_1^2=3, \quad y_1/y_2=\tfrac{1}{2}\approx y_1/y_\infty, \tag{5}$$

where y_2 is the depth after the weak hydraulic jump from the limiting depth y_1, while y_∞ is defined as the depth of the trailing edge below the original free water surface as shown in Fig. 2. In general, because of the loss of total head, $y_2<y_\infty$. The largest possible value of $y_2(\approx y_\infty)$ is provided by $M_1=\sqrt{3}$, that is, by Eq. (5). If $M_1>\sqrt{3}$, then y_2 is considerably decreased and the free stream water level behind the hydrofoil can force the hydraulic jump upstream until $M_1=\sqrt{3}$ and $y_2\rightarrow y_\infty$. However, this is only true for a hydrofoil in infinitely deep water since wave disturbances cannot be propagated upstream to the hydrofoil if $V_\infty>\sqrt{gh}$ (where h is the finite depth of the channel containing the hydrofoil). In this latter case it is not even necessary to have a hydraulic jump.

When hydrostatic conditions have been attained, the pressure coefficient may be written as

$$C_p=\frac{p-p_\infty}{\rho V_\infty^2/2}=\frac{2gy_\infty}{V_\infty^2}\left(\frac{y-y_\infty}{y_\infty}\right)=\frac{2}{M_\infty^2}\left(\frac{y}{y_\infty}-1\right), \tag{6}$$

V_∞ being defined as the velocity of the hydrofoil relative to the water. Then if y_∞ is again defined as the depth of the trailing edge below the original free water surface, Eqs. (5) and (6) give a limiting value of

$$C_{p_1}=-\frac{1}{M_\infty^2}=-\frac{gy_\infty}{V_\infty^2}. \tag{7}$$

This provides a rigorous limiting value since the actual depth $y_2\leqq y_\infty$ at the trailing edge, while the velocity there is always greater than V_∞, especially for thick hydrofoils or large angles of attack. In these latter cases, however, the vertical accelerations may be so large that hydrostatic conditions are never established and, therefore, Eq. (7) cannot provide a satisfactory limiting value.

Since hydrostatic conditions are never established near the leading edge, only near the critical depth which must occur aft of $C_p=0$, y_∞ must be defined at the trailing edge. Defining y_∞ in this manner also satisfies the Kutta-Joukowski condition, preventing an infinite viscous force at the sharp trailing edge by requiring that the pressure recovery be identical for the upper and lower surfaces. When $y_2=y_\infty$ Eq. (6) gives $C_p=0$ for the trailing edge, agreeing with the experimental data of Fig. 1 for the larger submergence depths. As the submergence depth decreases, $y_2<y_\infty$ and $C_p<0$, again in agreement with Fig. 1.

Figure 2 proves that Eq. (7) is in excellent agreement with the experimental data. Further justification of Eq. (7) and the assumption of the hydrostatic relations were obtained by measuring the water depths and finding them constant over the limiting flow portion of the hydrofoil, and agreeing with Eq. (5) after the jump.

Also shown in Fig. 2 is the absolute limit for the pressure coefficient following the hydrostatic assumption, when no jump is formed and the flow velocity increases indefinitely until y approaches zero so that Eq. (6) reduces to

$$C_{p_{y=0}}=-\frac{2}{M_\infty^2}=-\frac{2gy_\infty}{V_\infty^2}. \tag{8}$$

If Eq. (8) is exceeded, it proves that the hydrostatic relations are not attained and the vertical accelerations are not negligible.

The other critical line shown in Fig. 2 is the one that must be exceeded to produce air entrainment, provided the hydrostatic relation is applicable. This condition can be derived from the momentum change required to reduce the net pressure to atmospheric, so that

$$\int_0^y p\,dy-0=\rho g y^2/2=\rho V y(V_1-V),$$

or

$$y/y_1=1+\frac{1}{2M^2}, \tag{9}$$

since

$$yV=y_1V_1.$$

Then, if we further assume that the initial conditions are $y=y_\infty$, $M=M_\infty$ (not necessarily true, since hydrostatic conditions may not be reached ahead of the critical depth), Eqs. (6) and (9) yield

$$C_{p_{p=0}}=-\frac{2}{M_\infty^2}\left(\frac{1}{1+2M_\infty^2}\right). \tag{10}$$

[4] B. A. Bakhmeteff, *Hydraulics of Open Channels* (McGraw-Hill Book Company, Inc., New York, 1932), pp. 229, 249–252.

Even if Eq. (10) applied, and hydrostatic conditions held, it could be exceeded before serious air entrainment and flow separation occurred because the hydrofoil surface is very smooth and the curvature is small.

It is obvious that if hydrostatic flow conditions are not established, then Eqs. (7), (8), and (10) are not applicable, and in this case no limiting value can be predicted. This adverse condition would occur for very thick hydrofoils or large angles of attack because the severe vertical accelerations produced in these cases would preclude the possibility of ever establishing hydrostatic conditions. Another possibility invalidating Eq. (7) would be to have the average velocity correspond to $M_1>\sqrt{3}$ before the vertical acceleration became negligible and hydrostatic conditions were established. However, this seems to occur only with a hydrofoil in a finite depth channel when $V_\infty>\sqrt{gh}$ so no wave disturbances can be propagated upstream, and therefore Eq. (7) is not applicable, as previously pointed out.

It is easily shown that when $y_2=y_\infty$,

$$M_\infty^2=\frac{8M_1^2}{[(1+8M_1^2)^{\frac{1}{2}}-1]^3}, \tag{11}$$

which is of course the relation between the Froude number after the hydraulic jump (M_∞^2) to the upstream value (M_1^2). Equation (11) shows that

$$M_1^2\leqq 3 \quad \text{for} \quad M_\infty^2\geqq\tfrac{3}{8}. \tag{12}$$

This further proves the applicability of Eq. (7) as long as the hydrofoil moves in infinitely deep water, since Eq. (12) shows that values of $M_\infty^2>\frac{3}{8}$ cannot have $M_1>\sqrt{3}$ and still have $y_2\rightarrow y_\infty$.

It must be noted that the relations used to derive Eq. (7) are essentially for horizontal channels, since Bakhmeteff[4] has proved that the hydraulic jump relations are satisfactory for predicting results only in channels sloping downward less than 4 degrees. However, Eq. (7) itself may be expected to hold for greater angles than this, as shown in Fig. 2, because $(y_2-y_1)/y_1$ is practically independent of the channel slope (see p. 252 of reference 4). The limiting pressure coefficient only depends on this value which then may be taken as $\frac{1}{2}$ for all angles of attack between zero and 10 degrees. Visual observation made by the author indicated that Eq. (7) was definitely not applicable for angles of attack approaching 18 degrees. The indication was that channels sloping at 18 degrees or more actually could not support a weak hydraulic jump ($M_1=\sqrt{3}$) but could only maintain a very abrupt strong hydraulic jump. On the hydrofoil this strong jump would occur behind the trailing edge and produce an energy loss of over 50 percent of the total head. In this case, the Kutta-Joukowski condition can be violated; there is practically no lift but there is very large pressure drag, and Eq. (7) is of course not applicable.

Also, as is well known, small span hydrofoils approach a lift force at the surface equal to one-half the lift force at great depths; therefore, small span hydrofoils are not limited by Eq. (7). This is also obvious from a physical consideration of the span-wise flow which makes the hydrostatic assumption impossible for a small span (low aspect ratio) hydrofoil.

Fast Time Analysis of Intermittent Point-to-Plane Corona in Air. III. The Negative Point Trichel Pulse Corona*

M. R. AMIN†
Department of Physics, University of California, Berkeley, California
(Received September 21, 1953)

Fast oscilloscopic time analysis of the negative point Trichel pulse corona in room air at various pressures and gap geometries reveals the following data. The very short rise and quenching time of the pulse at atmospheric pressure observed by English is confirmed. Under these conditions the secondary action is a photoelectric liberation from the cathode and discharge extinguishes by dissociative attachment to give O^- ions. Decreasing pressures reduces space charge density, prolongs the discharge and brings in a secondary liberation by positive ion bombardment. Increasing potential at constant pressure at first decreases clearing time, leaving pulses unchanged and current increases proportional to repeat rate. At higher potentials, photon action is reduced relative to positive ion impact at the cathode, the discharge extends further into the gap, and quenching is very effective with less total ions reducing pulse size. With increasing repeat rate, current no longer increases proportional to repeat rate in consequence of smaller pulse height. The effects of increased point diameter are increased pulse size. Comparable clearing times for large and small points require higher potentials for the former. Thus, at constant pressure and potential, increase in point radius produces a decrease in frequency and increase in total charge per pulse. Conditioning of the point at high-current densities and repeat rates decreases the number of ions per pulse, which is of the order of 10^9 under normal circumstances, and shortens the pulse by reducing cathode work function and increasing photoelectric liberation, so that choking occurs with little ion action. Since repeat rate is slightly decreased by the charge liberation, the reduced pulse charge yields a reduced current through a conditioning that decreases work function. Comparison of ion movement across the gap during clearing time, with that for O_2^+ ions of known mobility in air, indicates that the rapid ion transit in Trichel pulses is caused by a majority of O^- ions with a mobility of about 4 cm²/volts second from near the cathode with some formation of O_2^- ions in transit.

As the potential of a negative point is increased, the threshold for a self-quenching intermittent discharge is reached where the current increases abruptly from 10^{-11} to about 10^{-9} to 10^{-7} amp. At the threshold the pulses are random and irregular in time, but with further increase of voltage they become quite regular in time. In fact they are so regular that they resemble relaxation oscillator pulses. Raising the potential more causes an increase in the current as well as the repeat rate of the pulses until a pulseless discharge is reached at about 10^6 cycles per second. Visually, the pulsating discharge appears as a typical glow discharge close to the point. There is a Crookes dark space, a bright negative glow, a Faraday dark space, and a fan-like positive column. Such a discharge generally appears in air or electron-attaching gases.

This type of intermittent discharge was first observed by M. O'Day, but was studied in detail by Trichel,[1] and interpreted by L. B. Loeb. A comprehensive analysis and discussion of the early work was presented in an article by L. B. Loeb[2] in 1948. For a recent systematic study of current-voltage characteristics in standard geometry, the reader is referred to the paper of H. W. Bandel.[3]

Except for the oscillogram of the photon pulse taken by W. N. English[4] for a single geometrical arrangement at atmospheric pressure, the data for time scale analysis, so far, have been inadequate. English's photomultiplier showed a surprisingly short time (4×10^{-8} second), for the duration of the Trichel pulse. The existing mechanisms could not account for this, until 1952, at which time L. B. Loeb[5] provided the explanation. According to this mechanism, the pulse builds up in 10^{-8} second by the formation of successive avalanches as the result of secondary electron emission from the cathode. Since the rise time of the pulse is so short, this secondary action can be accounted for only by photoelectric effect at the cathode. Using the newly observed cross sections for dissociative attachment of electrons to O_2 molecules forming O^- ions at about 3 ev of energy (Geballe and Harrison[6]), Loeb was able to explain the fast time of choking. Thus, the formation of O^- ions causes an accumulation of space charge *near* the point which quenches the pulse in some 2×10^{-8} second.

OBSERVATIONS‡

The equipment and techniques used were those indicated in an earlier paper.[7] As in the previous papers,[7,8] the reproducibility and generality of the results justify the presentation of only a few illustrative oscillograms.

Figure 1 trace *B* illustrates the short duration of the Trichel pulse at atmospheric pressures. The square

* This work was supported for one year on funds from the U.S. Office of Naval Research and for a second year under a fellowship from the Research Corporation of America.
† Now at the University of Oklahoma, Norman, Oklahoma.
[1] G. W. Trichel, Phys. Rev. **54**, 1078 (1938).
[2] L. B. Loeb, J. Appl. Phys. **19**, 882 (1948).
[3] H. W. Bandel, Phys. Rev. **84**, 92 (1951).
[4] W. N. English, Phys. Rev. **77**, 850 (1950).
[5] L. B. Loeb, Phys. Rev. **86**, 256 (1952).
[6] M. A. Harrison and R. Geballe, Phys. Rev. **84**, 1072 (1951).
‡ Oscillograms in Fig. 1 were obtained, using type 513*D* Tektronix. All other oscillograms were made from the traces of 511*D* Tektronix, and that is why the spikes in these pictures are broader than those shown in Fig. 1.
[7] M. R. Amin, J. Appl. Phys. **25**, 210 (1954).
[8] M. R. Amin, J. Appl. Phys. **25**, 358 (1954).

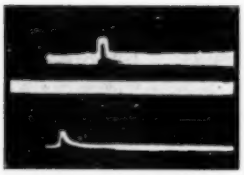


FIG. 1. Upper: 1×10^{-8} sec test pulse. Lower: Trichel photon pulse. 2.0-mm diameter point. 4-cm gap. Atmospheric pressure.

test pulse of 1×10^{-8} second, which was sent through the same system in Fig. 1, trace A, shows the short time resolution of the circuit. Figures 2, 3, and 4 show typical photon and electrical pulses as the different variables are changed. Figure 5a shows the regularity of Trichel pulses while Fig. 5b provides a good example for comparison of photon and electrical pulses.

There are certain very interesting features. At lower pressures and/or higher fields the initial spike is

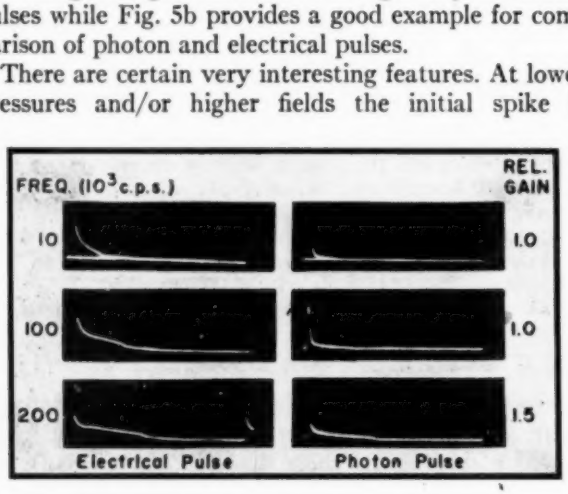


FIG. 2. Variation of Trichel pulse with repeat rate. 1.0-mm diameter point. 4-cm gap. Pressure = 100 mm Hg. $RC=10^{-8}$ sec. Sweep rate: 0.5 μsec/scale division.

followed by a plateau. This is seen in both the electrical and the photon pulses. The photomultiplier, seeing only the luminosity, does not record the movement of the ions in the high field region. The electrical pulse, however, records this motion. Therefore, the ratio of the height of plateau to its spike is less in the electrical pulse than in the photon pulse. Moreover, at the time

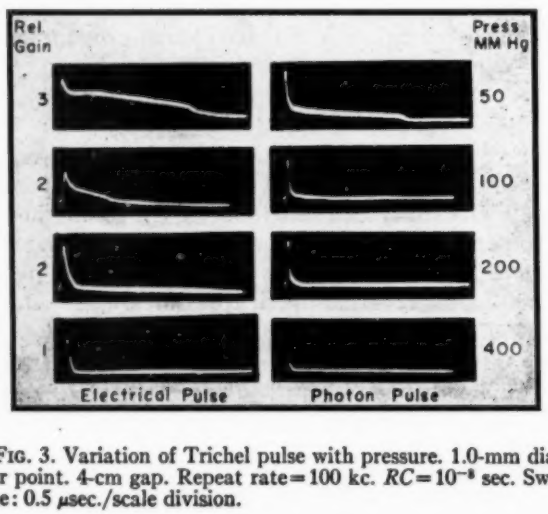


FIG. 3. Variation of Trichel pulse with pressure. 1.0-mm diameter point. 4-cm gap. Repeat rate = 100 kc. $RC=10^{-8}$ sec. Sweep rate: 0.5 μsec./scale division.

when the photon pulse drops rather sharply to zero, the electrical pulse shows merely a sudden break and with a change in slope gradually approaches the base line, Fig. 5b. This tail is a good evidence of ion motion in the high field region, since the electrons having a much higher mobility have been removed by this time. If this tail is extrapolated back to the earlier part of the discharge, then it appears that the relative shapes of electrical pulse caused by ionization resemble the photon pulse quite closely, Fig. 5b. Thus, ionization parallels excitation. In order to study the motion of negative ions in the low field region, reference is made to Fig. 6, where a large resistance ($R=10^6$ ohms) was used for this recording. For comparison the positive burst pulse is illustrated in the same figure.

Figure 7 shows the change in pulse shape as a result of the "Conditioning" (observed by English[9] and Bandel[3]) when the current reaches values near the transition to pulseless discharge. It is to be noted that the spikes are higher and the tails are shorter for a conditioned point. The conditioning also causes a

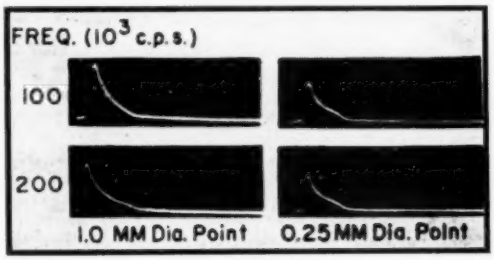


FIG. 4. Electrical Trichel pulses for two different sized points. 4-cm gap. Pressure = 200 mm Hg. $RC=10^{-8}$ sec. Sweep rate: 0.2 μsec/scale division. Left oscillogram: relative gain = 1.0. Right oscillogram: relative gain = 2.0.

change in frequency-current and frequency-voltage plots, Figs. 14 and 15.

The data on pulse frequency against current and frequency against potential are shown for several pressures and points in Figs. 8, 9, 10, 11, 12, and 13. Other information and characteristics like current against voltage and number of charges per pulse, can be obtained directly from these plots. Unfortunately, these plots cover a short range because of the following limitations.

1. Ultraviolet illumination of the point as a source of triggering electrons was not possible, since the ultraviolet source produces visible luminosity which gives spurious results and adds a high background noise to the photon pulse. Thus, studies could not be made at the Trichel pulse threshold. The appearance of multiple and diffusive spots at lower pressures and high potentials, and the limitation of voltage output of the power supply at higher pressures, prevented the study of higher frequencies.

2. At lower pressures (20 and 40 mm Hg for 0.5-mm

[9] W. N. English, Phys. Rev. 74, 170 (1948).

point) the frequency-current plots reached a maximum at about 130 microamperes, and then declined rapidly until the pulseless discharge was reached. In such cases the negative glow and other characteristics of the glow discharge disappeared, and a diffusive glow covered the whole point. Since such phenomena do not have the characteristics of Trichel pulses, their study was not pursued any further. Thus, the observation was limited to such pressures where the Crookes dark space, negative glow, etc., could be distinguished.

3. Since studies of Loeb indicate the fan of purple glow to result from movement of the steep potential front of the negative ion space charge that choked the discharge out into the gap after the luminous part of the glow is largely over, an attempt was made to observe this movement. However, the luminosity relative to the noise background was too weak to record and observe this movement.

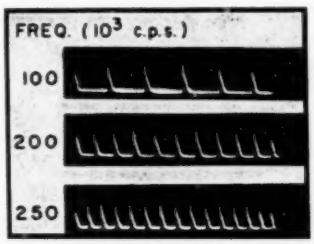


FIG. 5a. Regular electrical Trichel pulses. 1.0-mm diameter point. 4-cm gap. Pressure=100 mm Hg. $RC=10^{-8}$ sec. Sweep rate: 5 μsec/scale division. (Upper photographs.)

ANALYSIS AND INTERPRETATION OF THE OSCILLOGRAMS

First, an outline of the mechanism will be given and then it will be used to explain the interesting behavior of Trichel pulse as different variables change.

1. Mechanism

An electron emitted from the cathode under proper field conditions forms an avalanche. The photons created in this process strike the cathode liberating secondary electrons by photoelectric action (γp). The secondary electrons form a number of successive avalanches which give rise to a very rapidly growing pulse.[5,§] By dissociative attachment the outwardly moving electrons form a space charge of O^- ions very near the point which when sufficient chokes the pulse (Loeb mechanism).[5] If by the time the positive ions reach the cathode the field is still reduced, the secondary electrons created by positive ion bombardment (γi) are not capable of forming any more excitation and ionization. Thus, the discharge chokes completely until the space charge has cleared and another pulse can start. On the other hand, if upon the arrival of positive ions at the

§ The positive ions formed by the avalanches are created in a dense cloud at some distance from the point. These ions move into the cathode slowly.

FIG. 5b. Electrical and photon Trichel pulses. 1.0-mm diameter point. 4-cm gap. Pressure=50 mm Hg. $RC=10^{-8}$ sec. Sweep rate=0.5 μsec/scale division. (Lower photographs.)

point, the reduced field has not developed because of insufficient density of electron attachment (lower pressures and/or higher fields) the secondary electrons liberated are capable of forming more avalanches, and the discharge continues by γi action until enough space is formed to choke the pulse completely. It must be pointed out here that when the positive ions become active in secondary electron emission, the dicharge will be restricted to a narrow and dense region along the axial direction in comparison to the discharge with photon action. Such a constriction of the discharge is to

FIG. 6. Electrically induced pulses. 1.0-mm diameter point. 4-cm gap. Atmospheric pressure. $R=10^6$ ohms. $RC=10^{-4}$ sec. Sweep rate: 1000 μsec/scale division. Upper: burst pulse at 7.6 kv. Lower: Trichel pulse at 8.2 kv.

be expected, since the field is highest along the axis, and positive ions are not widely dispersed as are the photons.

From the foregoing discussion, therefore, one would expect the space charge formed by γp action to be more nearly in the form of a hemispherical shell (because of the diffusive photon action) and the space charge accumulated by the γi action (because of constricted discharge) to be in a narrow and dense region in the direction of axis.

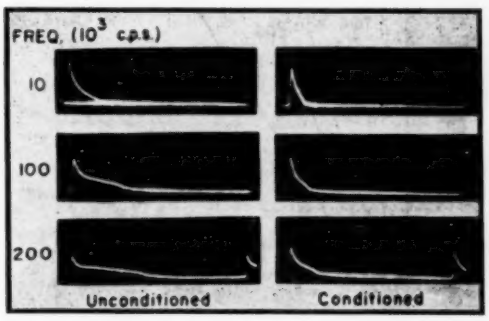


FIG. 7. Electrical Trichel pulses. 1.0-mm diameter point. 4-cm gap. Pressure=100 mm Hg. $RC=10^{-8}$ sec. Sweep rate: 0.5 μsec/scale division. Left oscillogram: relative gain=1.5. Right oscillogram: relative gain=1.0.

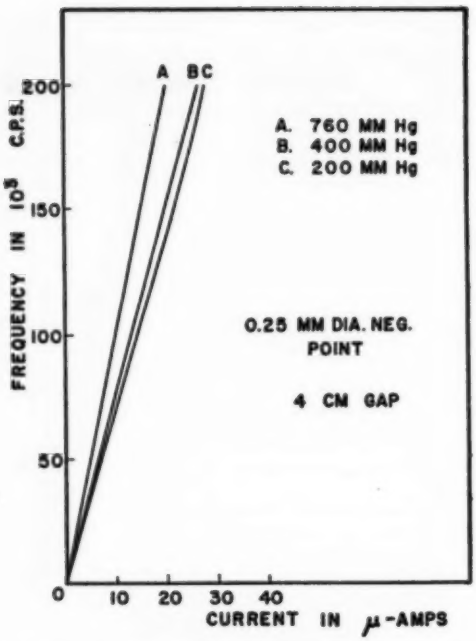


FIG. 8. Frequency *versus* current for Trichel pulse corona.

2. The Effect of Changing the Pressure

At high pressures the X/p (ratio of field to pressure) at onset is slightly lower than at lower pressures. The mean free paths are short. With high pressures at low X/p, excitation is high, yielding dense photon liberation near the surface. The dense photon action produces a rather heavy burst of current through a photoelectric (γp) secondary mechanism.[5] At high pressures the dissociative attachment occurs in a relatively short distance beyond the positive ion space charge cloud producing a negative ion space charge of high density close to the point. The field from this rapidly quenches the discharge. Thus, the pulse rises very sharply and declines rapidly, Fig. 1 trace b.

Decreasing the pressure results in diffusion of photon production, a reduction of photon density, and, therefore, a decrease in the number of photon striking the cathode. Decrease of the molecular density reduces the density of the negative ion space charge formation by dissociative attachment, and projects it further into the gap. This requires that more ions be created to yield fields that quench the discharge. Thus, the number of ions per pulse will increase, Figs. 8–10. The rate of ion production being decreased and the required ions for quenching being increased, a longer time for the buildup of space charge for quenching is necessary. In this longer time the positive ions reach the cathode in appreciable number. The space charge is then provided by the ionization resulting from the very efficient secondary electron emission as a result of positive ion impact at the cathode. The effectiveness of γi action at lower pressures is also increased by the higher positive ion energies and drift velocities.

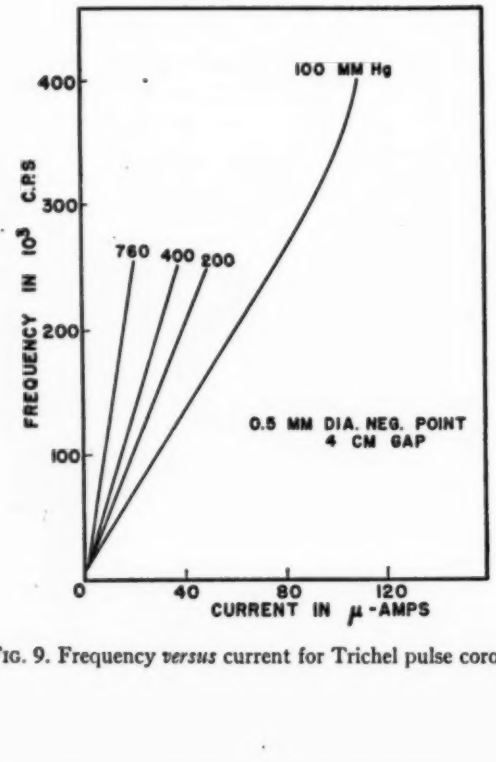


FIG. 9. Frequency *versus* current for Trichel pulse corona.

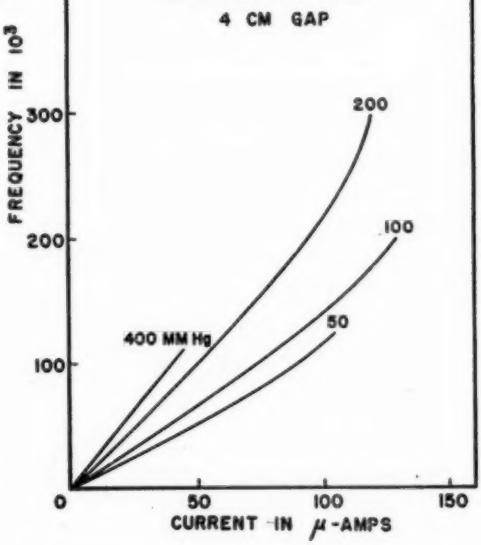


FIG. 10. Frequency *versus* current for Trichel pulse corona.

The effect of these conditions is clearly shown in the oscillograms of Fig. 3. Thus, with decreasing pressure, reduction in the height of the spike (photoelectric action γp), and an increase in duration of plateau (positive ion action γi) are observed. Obviously, the clearing time between pulses will decrease relative to those at higher pressures under equivalent condition, because of the higher negative ion mobility. Repeat rates will thus be higher for equivalent potentials, Figs. 11–13.

3. The Effect of Changing the Potential

As the potential is increased beyond the threshold, first the frequency rises linearly with current and then the current grows more slowly than the repeat rate. This indicates a constant total number of charge per pulse for the linear region and a decrease in total charge per pulse at higher currents and voltages, Figs. 8–10.

It appears that the increase in potential above threshold does not affect the fields across the active discharge region. This lies in the inhibiting effect of the space charge of the negative ions. Increasing potential merely acts to clear the negative ion space charge more rapidly. Therefore, the frequency and current rise nearly linearly with potential (Ohm's law regime). In this region only the γp action is effective and when the positive ions arrive at the cathode, the field is still

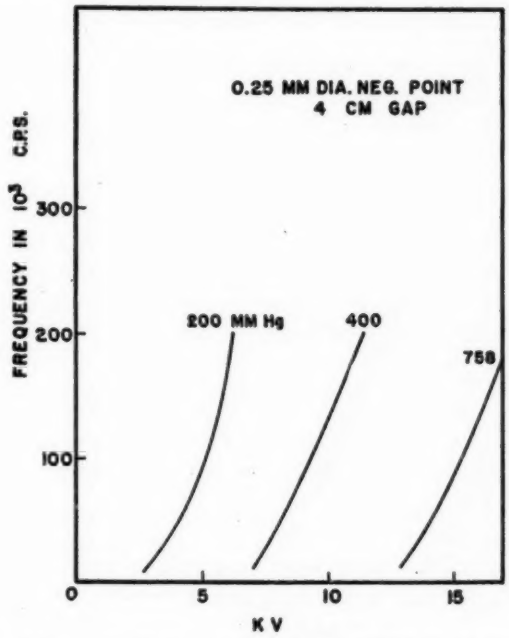


Fig. 11. Frequency *versus* potential for Trichel pulse corona.

quite low owing to the presence of negative space charge. Thus, the positive ion secondary action is ineffective.

With further increase in voltage, marked changes in mechanism take place since the increase in potential affects the active discharge region. As the result of increase in field, the negative ion space charge is formed further away from the point. The positive ions enhance the fields of the cathode. The ion mobilities are higher, and, thus, secondary electron liberation by positive ion bombardment at cathode is more intense and begins earlier and lasts longer. With effective ionization further out in the gap, at higher fields photons are lost to the cathode by diffusion. In consequence, the pulses will have a longer plateau and a smaller spike, Fig. 2. The reduction in the height of the spike is due to a decrease in γp action, and the elongation of the tail is

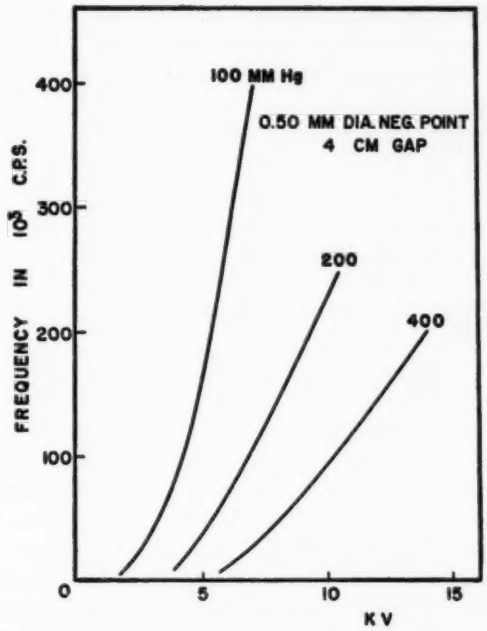


Fig. 12. Frequency *versus* potential for Trichel pulse corona.

due to the fact that a longer time is required for quenching the discharge.

As was mentioned in the discussion of the effects of lowered pressure, whenever the positive ions are active, the discharge is restricted to a narrow and dense region along the axis; that is, increased in length along the axis but constricted radially. In the restricted volume of ionization, the space charge density of negative ions formed is great enough to quench the discharge with a smaller total number of negative ions than under conditions where the discharge spreads

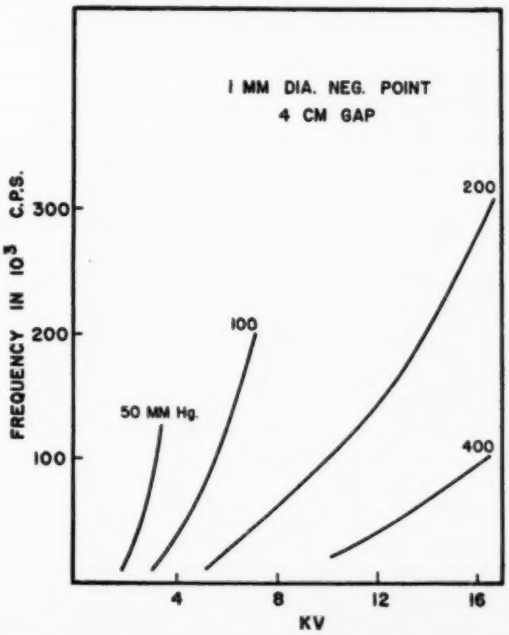


Fig. 13. Frequency *versus* potential for Trichel pulse corona.

radially over a greater area of the point, even though quenching requires a longer time. Thus, the decrease in the total charge per pulse is explained, although γi action is enhanced.

Increase in frequency of pulses with increase in potential is expected since it is predominantly the dead time,‖ or time of adequate removal of the space charge, and *not* the time of quenching which determines the frequency. When the dead time becomes very short, so that the space charge is removed as fast as it is formed, a pulseless Townsend-type discharge is achieved.

4. The Effect of Changing the Geometry

As has been pointed out by Trichel[1] and as was observed here, the variation of gap length is not a significant factor in these studies. However, the change in point radius is of interest, Fig. 4. The more efficient γp action for the larger point (higher spike) is to be expected, since it has a larger area for emission of secondary electrons. Also, the strong field extends further into the gap and more ions are needed to choke the pulse. In order to obtain the same frequency at a given pressure the larger point would require a higher voltage, since the dead time is dependent on the value of X/p at the point; hence, keeping the potential and pressure constant, an increase in point radius causes a decrease in frequency and an increase in total charge per pulse.

5. The Effects of "Conditioning"

Running the discharge for some time at higher frequencies and/or extending potentials to the region of the pulseless discharge causes some changes in the characteristics of Trichel pulse corona. Such changes were reported by English,[9] Bandel,[3] and C. G. Miller,[10] but were not understood. Figure 7 illustrates this effect. It has been termed "conditioning" in the past without knowing exactly what it implied. It must be pointed out, however, that the conditioning was not due to a change of the geometry of the point, since after a day or so the point nearly recovered its original characteristics.

The increase in spike height and reduction in plateau and total duration here observed, Fig. 7, at once permit an explanation of the conditioning process. It is clear that the increased spike implies a very great increase of γp, the photoelectric secondary emission. In consequence, as at high pressures, the very heavy current quickly produces enough negative ions near the point to choke off the discharge rapidly. That is, a high γp yields such a current that before many positive ions have moved in to give a γi plateau, the discharge is quenched. This can only happen if the work function of the cathode has decreased, which can yield a very large increase in γp. The rapid choking by heavy current increases the space charge close to the point in a short time, so that negative ions accumulate faster than they are drawn away. In comparison to a slower buildup of the pulse, this means that less total ions are needed to choke the pulse. Thus, the number of ions per pulse is decreased when conditioning occurs, Fig. 14.

Since the photon-conditioned γp predominates as a source of current, and by its nature produces the negative space charge close to the point and radially spreads over it shortening quenching time, it will also increase clearing time slightly. That is, in contrast to the removal of longer column of choking negative space charge produced by γi, the ion distribution around the point with γp action requires somewhat more time for removal. It should be pointed out that English, Bandel, and Miller had remarked on the changing of negative glow from a broader spot to a narrower, deeper spike under conditions where we now know γi to have been active. Thus, the conclusions drawn are in conformity with visual manifestations. In consequence of conditioning with decreased work function of the cathode, the mechanism is altered, spreading the discharge radially, increasing photon action, decreasing γi action, decreasing quenching time and charge per pulse, and decreasing frequency at a fixed potential as seen in Fig. 15.

6. Comparison of Trichel Pulse with Burst Pulse

Using a large resistor in series with plane enables one to observe the motion of ions (positive or negative)

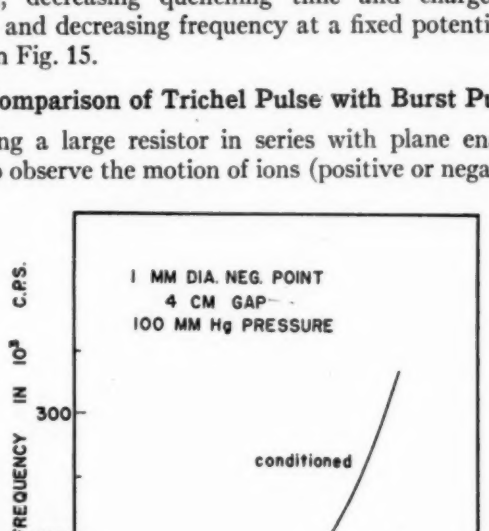


FIG. 14. Frequency *versus* current as observed where conditioning occurs.

‖ The dead time here is the time required to remove the space charge far enough from the point so that another pulse can start. It is essential to distinguish the *dead time-like* intervals for movement of local space charges far enough from the point to permit new pulse formation from the *clearing time-like* intervals involving clearing of space charge more or less completely from the gap.

[10] C. G. Miller and L. B. Loeb, J. Appl. Phys. 22, 614 (1951).

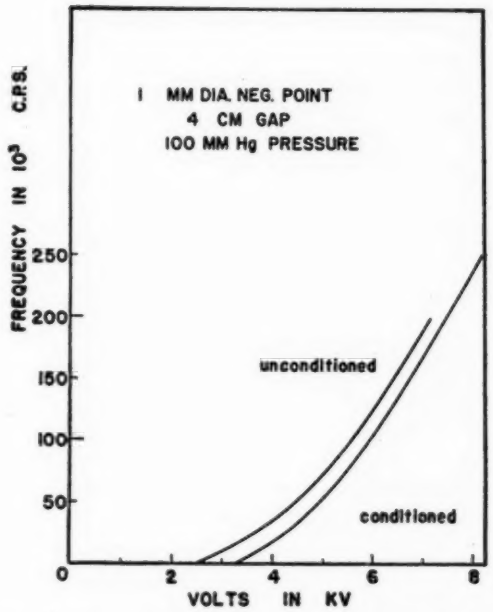


FIG. 15. Frequency *versus* potential as observed where conditioning occurs.

across the gap. Figure 6 illustrates such observations for the positive burst pulse trace *a* and negative Trichel pulse trace *b* under exactly the same condition, except that the potential for burst pulse was somewhat less than that for the Trichel pulse. The inductive kick signifies the time of occurrence of the pulse. This spike is followed by a plateau indicative of the motion of the ions in the low field region. The motion of the ions in the high field region is not registered because of the long *RC* time constant of the circuit ($RC = 10^{-4}$ second). From the comparative study of the two types of pulses, the following interpretations were made.

It is to be noted that the positive ions of the burst pulse (time of formation 5×10^{-4} second) have a time of crossing of about 5000 microseconds. This is indicated by the break in the plateau of the upper trace. Following the break, the trace approaches the zero position gradually in some 10^{-3} second. This is because of the long time of formation of the burst pulse (5×10^{-4} second) and diffusion of ions. The positive ions are identified as O_2^+ ions, since the work of R. N. Varney[11] has shown that the O^+ ions, having a higher ionization energy than O_2^+ ions, cannot exist under the conditions of this experiment, and lose their charge to form O_2^+ ions by exchange forces. According to Varney, the mobility of O_2^+ ions in oxygen at atmospheric pressure is 2.25 cm²/volt second. The mobility of O_2^+ ions in air so far has not been measured accurately. However, from the above value and the early investigations[12] on mobilities of positive ions in air, the mobility of O_2^+ ions in air at atmospheric pressure is expected to be between 1.8 and 2.2 cm²/volt second.

A glance at the Trichel pulse reveals that the break appears some 2500 microseconds after the time of formation of the pulse. Since the mobility of O_2^- ions in air at atmospheric pressure, according to Doehring,[13] is 2.68 cm²/volt×second, the break could not be accounted for by the arrival of O_2^- ions at the plate if these ions are formed in a region only a few millimeters from the point. This indicates the presence of ions of higher mobilities. Therefore, on the basis of our mechanism, it is logical to asume that the break is indicative of the arrival of O^- ions. Such an assumption would mean that the O^- ions in air have a mobility of about 4 cm²/volt second. The broad tail following the break and gradually approaching zero is explained in the following manner.

According to the mechanism some of the electrons which do not undergo dissociative attachment, say some 20 percent at 760 mm Hg, form O_2^- ions by Bradbury-Bloch[14] process in the lower field region. Thus, a very diffuse cloud of O_2^- ions is formed in a relatively large distance from the point, say, a few millimeters. This could account for the long broad tail. Also, the self-repulsion and diffusion of a dense cloud of O^- ions contributes to this broadening appreciably. It must be pointed out that if there were present *only one type of ion* (created near the point), the break would be much sharper and the tail much shorter, since the time of formation of Trichel pulse (10^{-8} second) is about 10^4 times shorter than that of the burst pulse. Yet the burst pulse gives a comparatively longer plateau and shorter tail. It is true that the field is somewhat higher for the Trichel pulse than for the burst pulse. However, the difference in field would be effective only for motion in the high field region, but the small difference would not affect appreciably the time of crossing the whole gap.

The following explanation cannot be ignored entirely. If the O^- ion formed complex ions of lower mobility than that of O_2^+ ions in the lower field region, then it is possible to account for the break at 2500 microseconds as being caused by O_2^- ions that form $\frac{2}{3}$ the way to the anode and the long tail as being caused by the clustered O^- ions. Since for choking the pulse, O^- ions must form near the point to the extent of at least 50 percent of the electrons, it is clear that the rather sharp break at 2500 microseconds with the very gradual tailing off of the pulse is inconsistent with the gradual formation of O_2^- ions in mid-gap which would give a very broad first break.

The writer wishes to express his gratitude to Professor L. B. Loeb, under whose suggestion and direction this work was done, for his guidance and valuable assistance throughout the work.

[11] R. N. Varney, Phys. Rev. **89**, 708 (1953).
[12] L. B. Loeb, *Fundamental Processes of Electrical Discharges in Gases* (John Wiley and Sons, Inc., New York, 1939), pp. 1–31.
[13] Doehring, Z. Naturforsch. **7a**, 253 (1952).
[14] F. Bloch and N. E. Bradbury, Phys. Rev. **38**, 689 (1935).

Volume and Surface Recombination Rates for Injected Carriers in Germanium

J. P. McKelvey and R. L. Longini
Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania
(Received September 28, 1953)

A study of bulk minority carrier lifetimes and surface recombination rates has been made, using experimental equipment similar to that of Navon, Bray, and Fan, wherein the time-decay of carriers injected electrically by a pulse is observed. The diffusion theory for carriers of bulk lifetime τ has been worked out, considering a definite probability of nonrecombination, or reflection coefficient, to be associated with the surfaces of the sample. A simple relationship between this quantity and the surface recombination velocity has been derived. If the surface recombination velocity is known, the effective trapping area at the surface may be calculated from this relation. Values of bulk lifetime and surface recombination velocity have been found for both n- and p-type germanium samples, and for a wide variety of surface treatments. By suitable modifications in experimental technique, this experiment may be utilized to measure drift mobility of minority carriers.

INTRODUCTION

An experimental technique developed by Navon, Bray, and Fan[1] has been used in this laboratory to make an extensive study of bulk lifetime and surface recombination in germanium. It has been found possible to utilize diffusion theory appropriate to the sample geometry used in the experiments to separate bulk and surface effects, and, by slight modifications in experimental technique, to calculate drift mobility of minority carriers from recombination measurements. Briefly, the experimental technique involves electrical injection of minority carriers into a sample of germanium in the form of a rectangular parallelepiped by the application of a voltage pulse across the sample, and observation of the decay of the injected carrier density as a function of time thereafter by measuring the conductivity of the sample with a very small sweep voltage. A schematic diagram of the experimental apparatus is shown in Fig. 1.

DIFFUSION AND RECOMBINATION OF INJECTED CARRIERS

Carriers injected into a semiconductor move by the usual processes of diffusion, and the injected carrier density, that is, the actual carrier density minus the thermal equilibrium carrier density, satisfies the diffusion equation (1).

$$D(\nabla^2 n)-\frac{n}{\tau}=\frac{\partial n}{\partial t}. \tag{1}$$

Here n is the injected carrier density, D is the diffusion coefficient for the injected carriers, and τ is the average lifetime of injected carriers in the bulk of the material. It will be assumed that the diffusion equation holds in a rectangular solid whose thickness, width, and length are given by x_0, y_0, and z_0, respectively. It will later be necessary to assume that $x_0 \ll y_0 < z_0$. In this geometry we express n as $n(x,y,z,t)$, and it is assumed that for $t=0$ there is present a uniform injected carrier density n_0 throughout this volume, that is,

$$n(x,y,z,0)=n_0=\text{const.} \tag{2}$$

It is necessary also to account for the fact that trapping and subsequent recombination at the surface of the sample occurs more or less independently of the bulk recombination. The surface recombination may be characterized by the probability of nonrecombination for a carrier upon colliding with the surface potential barrier. This quantity is termed the surface "reflection coefficient" for carriers and denoted by R. Figure 2 represents a situation where A carriers per unit time are introduced into the surface region. Of this number RA are "reflected" back to the bulk. Those entering the bulk are subject to various scattering and absorption processes, so that a fraction B will eventually return to the surface. The carriers are thus subject to a "multiple reflection" process, R being the reflection coefficient of the surface and B being the effective reflection coefficient of the bulk, as shown in Fig. 2. Summing carrier flux to the right and to the left, one obtains (3) and (4), respectively.

$$F_1=A\sum_{i=0}^{\infty}R^iB^i=\frac{A}{1-RB} \tag{3}$$

$$F_2=AR\sum_{i=0}^{\infty}R^iB^i=\frac{AR}{1-RB}. \tag{4}$$

It must be true that if $R=1$

$$F_1=F_2=\int_0^\infty v_x\Phi(v_x)dv_x=\frac{n_s}{\sqrt{\pi}}\left(\frac{m}{2kT}\right)^{\frac{1}{2}}\int_0^\infty v_x \times\exp[-mv_x^2/2kT]dv_x=n_s\left(\frac{kT}{2\pi m}\right)^{\frac{1}{2}}, \tag{5}$$

where $\Phi(v_x)$ is the Boltzmann velocity-in-x distribution and n_s is the injected carrier concentration in the surface layer. Since the mean thermal velocity is given by

[1] Navon, Bray, and Fan, Proc. Inst. Radio Engrs. 40, 1342 (1952).

$\bar{c}=[(8/\pi)(kT/m)]^{\frac{1}{2}}$, one may write (5) in the form

$$F_1=F_2=\frac{n_s\bar{c}}{4}. \tag{6}$$

In these equations, m refers to the effective mass of carriers in the semiconductor.

If $R<1$, there will be a net flow of carriers toward the surface, and F_1 will be greater than F_2. In this case, if the fractional concentration variation over the carrier mean free path is small, one may consider that the flux F_1 which arrives at the surface originates at a virtual source situated to the left of the surface (Fig. 2) a distance $\alpha\lambda$, where λ is the carrier mean free path and α is a numerical constant of the order of unity. Similarly, F_2 may be assumed to originate at a virtual source situated a distance $\alpha\lambda$ to the right of the surface. Then, substituting the value of n for these virtual sources in (6),

$$F_1=\left[n_s-\alpha\lambda\left(\frac{\partial n}{\partial x}\right)_s\right]\frac{\bar{c}}{4}, \tag{7}$$

$$F_2=\left[n_s+\alpha\lambda\left(\frac{\partial n}{\partial x}\right)_s\right]\frac{\bar{c}}{4}. \tag{8}$$

Adding (3) and (4), equating the result to the sum of (7) and (8), and solving for A, the result is

$$A=\frac{n_s\bar{c}}{2}\frac{1-RB}{1-R}. \tag{9}$$

Subtracting (4) from (3) and utilizing (9), one obtains (10).

$$F_1-F_2=\frac{n_s\bar{c}}{2}\frac{1-R}{1+R}. \tag{10}$$

Here, F_1-F_2 is the net flow toward the surface, which must equal the normal component of the classical diffusion flow, $-D(\mathbf{U}\cdot\nabla n)$, at the surface, and we shall identify the coefficient of n_s in (10) with the parameter s, usually referred to as the surface recombination velocity. $\mathbf{U}$ denotes an outwardly directed unit vector normal to the surface. Thus

$$F_1-F_2=-D(\mathbf{U}\cdot\nabla n)_s=\frac{n_s\bar{c}}{2}\left(\frac{1-R}{1+R}\right)\equiv n_s s. \tag{11}$$

This immediately establishes a further boundary condition on (1):

$$-D(\mathbf{U}\cdot\nabla n)_s=n_s s \tag{12}$$

and also expresses the surface recombination velocity in terms of a more fundamental quantity, the probability R of nonrecombination for a collision with the surface. Using (11), (7), and (8),

$$s=\frac{\bar{c}}{2}\frac{1-R}{1+R}=\frac{F_1-F_2}{n_s}=-\frac{\alpha\lambda\bar{c}}{2}\left(\frac{\partial \ln n}{\partial x}\right)_s. \tag{13}$$

From (7), (8), and (11), noting that $D=\lambda\bar{c}/3$, it is an easy matter to show that in fact $\alpha=\frac{2}{3}$.

It is necessary then to solve the diffusion equation (1) subject to the boundary conditions (2) and (12). The equation can be solved in a rectangular solid of dimensions x_0, y_0, z_0, but the solution is extremely complicated. We shall assume that x_0 is small compared to y_0 and z_0, obtain the solution to the diffusion equation in a slab of thickness x_0 extending to infinity in the y and z directions, as a first approximation, and draw upon the results of the 3-dimensional solution to make correc-

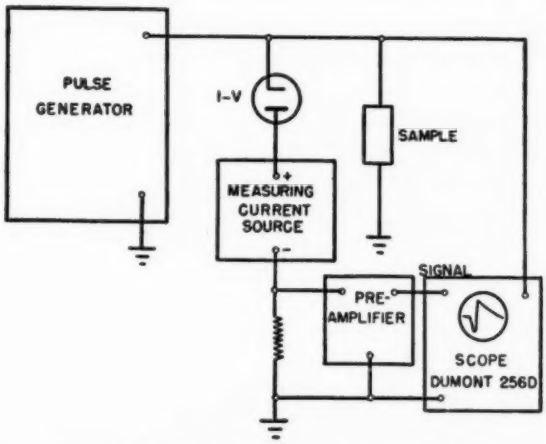


FIG. 1. Schematic block diagram of lifetime measuring apparatus.

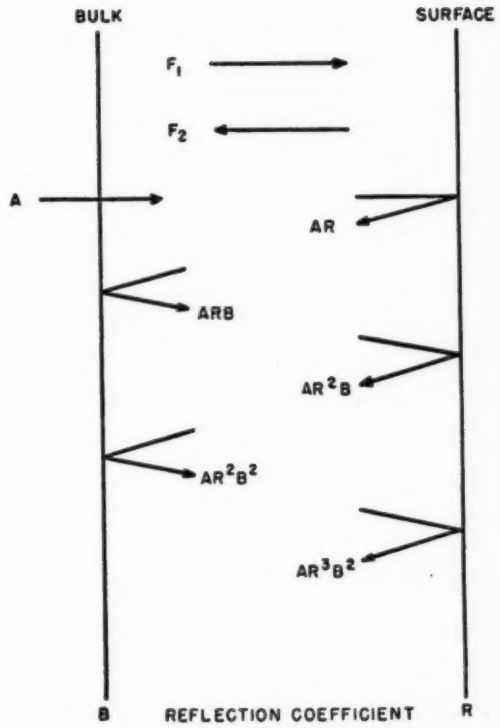


FIG. 2. Multiple reflection representation of carrier diffusion in the neighborhood of a surface.

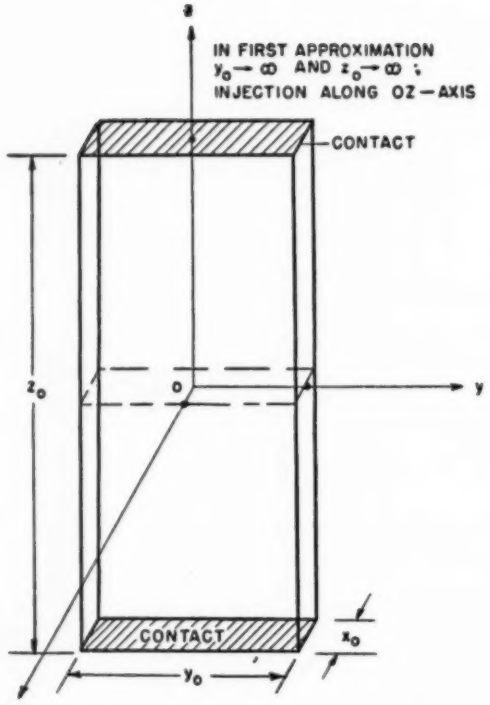


FIG. 3. Assumed orientation of sample for diffusion calculations. As a first approximation y_0 and z_0 are assumed to be infinite, resulting in a large slab of material of thickness x_0.

tions to the first approximation for side and end effects arising from the fact that y_0 and z_0 are finite.

Proceeding in this manner, the boundary value problem becomes

$$D\frac{\partial^2 n}{\partial x^2}-\frac{n}{\tau}=\frac{\partial n}{\partial t} \tag{14}$$

$$n(x,0)=n_0; \qquad D\left|\frac{\partial n}{\partial x}\right|_s=n_s s. \tag{15}$$

The equation may be solved under these boundary conditions by the classical methods of separation of variables and expansion of the initial carrier density in terms of eigenfunctions which satisfy the boundary conditions.[2] If the origin is taken in the center of the sample (Fig. 3), the solution of (14) which satisfies the boundary conditions (15) is

$$n(x,t)=4n_0\varepsilon^{-t/\tau}\sum_{i=1}^{\infty}\frac{\sin(\xi_i\pi/2)}{\xi_i\pi+\sin\xi_i\pi}\times\exp[-\pi^2\xi_i^2Dt/x_0^2]\cos(\xi_i\pi x/x_0), \tag{16}$$

where the set of eigenvalues (ξ_i) is chosen such that the eigenvalue ξ_i is the ith positive root of the equation

$$\operatorname{ctn}(\xi\pi/2)-\xi\pi D/sx_0=0. \tag{17}$$

[2] See for instance W. Shockley, *Electrons and Holes in Semiconductors* (D. Van Nostrand Company, Inc., New York, 1950), p. 319.

The actual conductivity change $\delta\sigma$ is easily found from (16), since

$$\delta\sigma=\frac{e_0\mu}{x_0}\int_{-x_0/2}^{x_0/2}n(x,t)dx=8n_0e_0\mu e^{-t/\tau}\times\sum_{i=1}^{\infty}\frac{\sin^2(\xi_i\pi/2)}{\xi_i\pi(\xi_i\pi+\sin\xi_i\pi)}\exp[-\pi^2\xi_i^2Dt/x_0^2], \tag{18}$$

where e_0 is the electronic charge and μ the drift mobility of the minority carriers. (18) represents the time decay of the injected carrier density and corresponds to the trace observed on the screen of the oscilloscope. The surface recombination velocity influences the values of the eigenvalues (ξ_i) through (17) and thus influences the time decay of (18).

It is desirable to analyze (18) as a function of time in such a way as to be able to infer the surface recombination velocity, assuming the bulk lifetime to be known. This can be accomplished by observing that whatever the value of s may be, the value of ξ_1 (and thus the value of s) may be obtained by measuring the decay constant of (18) after some time has elapsed after injection. This is true because after a comparatively short time the magnitudes of all the terms in the summation in (18), other than the first, become negligible compared to that of the first term. For small values of $s(s\ll 2D/x_0)$, from (17) it is clear that $\xi_i\cong 2i-2$, and under these conditions, using L'Hôpital's rule to evaluate the limits problem which arises when $i=1$, it can be shown that the amplitude factor of the terms for which $i>1$ is small compared to the amplitude factor of the first term. Also, the time constants of the higher terms are smaller and thus after a short time these terms become very small indeed. For large values of $s(s\gg 2D/x_0)$, (17) shows that $\xi_i\cong 2i-1$. The time constant of the second term is only about $\frac{1}{9}$ that of the first, and the amplitude factor of the second term is also about $\frac{1}{9}$ that of the first, so again after a short time this term can be neglected. The third and higher terms are even smaller. For intermediate values of s, both of these arguments apply to a certain extent, the result being essentially the same; one may measure ξ_1 by measuring the time constant of the "tail" of the function (18).

Neglecting terms in the summation in (18) other than the first, and denoting the over-all observed decay constant as τ_0, we may write

$$\frac{1}{\tau_0}=\frac{1}{\tau}+\frac{\xi_1^2\pi^2D}{x_0^2}. \tag{19}$$

Solving this equation for ξ_1, substituting in (17), and solving for s, the result is

$$s=\left[D\left(\frac{1}{\tau_0}-\frac{1}{\tau}\right)\right]^{\frac{1}{2}}\tan\left\{\frac{x_0}{2}\left[\frac{1}{D}\left(\frac{1}{\tau_0}-\frac{1}{\tau}\right)\right]^{\frac{1}{2}}\right\}. \tag{20}$$

This equation has been derived under the assumption of surface recombination in an infinite slab of thickness x_0. To deal rigorously with recombination in a rectangular solid, it is necessary to solve a 3-dimensional boundary value problem (1) using boundary conditions (2) and (12). It is possible to avoid the complication involved in using directly the results of this solution by incorporating the effects of recombination on the side and end surfaces of the sample into a corrected sample thickness, or "virtual thickness," $\bar{x}_0(\leqslant x_0)$. This corrected thickness $\bar{x}_0$ is defined as the thickness a sample of infinite extent in the y and z directions would have to have to produce the same observed decay constant τ_0, which the actual sample under consideration, with finite dimensions x_0, y_0, z_0, exhibits. Such a treatment is an approximation and demands that x_0 be small compared to y_0 and z_0. (See Fig. 3.) In actual practice it turns out that $x_0 \sim y_0/3$ is satisfactory.

The solution of the diffusion equation in a 3-dimensional rectangular solid under the boundary conditions (2) and (12) is a triple summation of products of functions of the same form as those shown in the summation in (18). These functions involve three sets of eigenvalues, (ξ_i, η_j, ζ_k) for the (i,j,k) term of the triple summation corresponding to the x,y, and z directions, respectively. It can be shown in a manner similar to that used in the first approximation that the time constant of the "tail" of the $\delta\sigma$ *vs* t curve obtained from the rigorous solution of the diffusion equation is essentially that of the (1,1,1) term of the triple summation, the effects of the higher terms becoming small after a comparatively short time following injection. The analogs of Eqs. (19) and (17), respectively, are

$$\frac{1}{\tau_0}=\frac{1}{\tau}+\pi^2 D\left(\frac{\xi_1^2}{x_0^2}+\frac{\eta_1^2}{y_0^2}+\frac{\zeta_1^2}{z_0^2}\right) \tag{21}$$

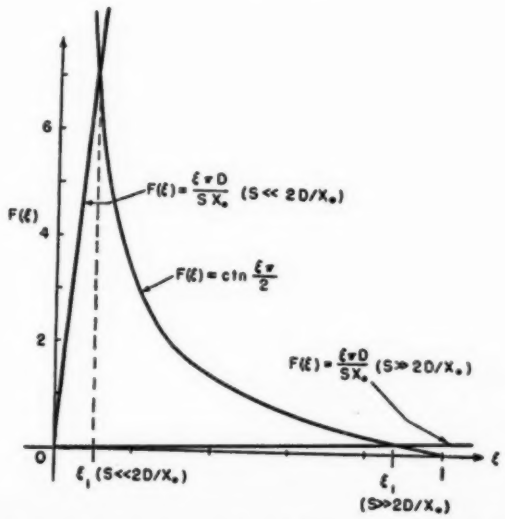


FIG. 4. Graphical representation of the roots of an equation of the form of (22) for large and small values of s.

and

$$\left.\begin{aligned}\operatorname{ctn}(\xi_1\pi/2)&=\xi_1\pi D/sx_0\\ \operatorname{ctn}(\eta_1\pi/2)&=\eta_1\pi D/sy_0\\ \operatorname{ctn}(\zeta_1\pi/2)&=\zeta_1\pi D/sz_0\end{aligned}\right\}. \tag{22}$$

In this case ξ_1, η_1, and ζ_1 cannot be eliminated between the four equations (21) and (22) to give an explicit formula for s, analogous to (20). Nevertheless it is possible to calculate s by making suitable approximations and introducing the virtual thickness $\bar{x}_0$.

Suppose that s is small ($s \ll 2D/z_0$). Then, in (22) the cotangents will be much larger than their arguments, which implies that the eigenvalues (ξ_1, η_1, ζ_1) will be small compared to unity. This situation is shown graphically in Fig. 4. In this case, one may assume in (22) that ctn $x \cong 1/x - x/3$ and eliminate ξ_1, η_1, ζ_1 between the four equations. The result is

$$\frac{1}{\tau_0}-\frac{1}{\tau}=12sD\left[\frac{1}{x_0(6D+sx_0)}+\frac{1}{y_0(6D+sy_0)}+\frac{1}{z_0(6D+sz_0)}\right], \tag{23}$$

which reduces to

$$\frac{1}{\tau_0}-\frac{1}{\tau}=2s\left(\frac{1}{x_0}+\frac{1}{y_0}+\frac{1}{z_0}\right),$$

if sx_0, sy_0, and sz_0 are neglected compared to $6D$. We shall write the above equation

$$\frac{1}{\tau_0}-\frac{1}{\tau}=\frac{2s}{x_0}(1+\alpha+\beta), \tag{24}$$

where

$$\alpha=\frac{x_0}{y_0} \qquad \beta=\frac{x_0}{z_0}. \tag{25}$$

In the first approximation the analog of this equation is (19). In (17) if $s \ll 2D/x_0$, ctn $x \cong 1/x$ and $\xi_1^2\pi^2 = 2sx_0/D$, whence (19) can be written

$$\frac{1}{\tau_0}-\frac{1}{\tau}=\frac{2s}{x_0}. \tag{26}$$

Comparing (24) and (26), it is obvious that if τ is identical for the actual finite sample and for a hypothetical one-dimensional sample of thickness $\bar{x}_0$, then

$$\bar{x}_0=\frac{x}{1+\alpha+\beta} \qquad (s \ll 2D/z_0). \tag{27}$$

If, on the other hand, $s \gg 2D/x_0$, then the eigenvalues (ξ_1, η_1, ζ_1) are all nearly unity (see Fig. 4). In this case one can write (21)

$$\frac{1}{\tau_0}-\frac{1}{\tau}=\pi^2 D\left(\frac{1}{x_0^2}+\frac{1}{y_0^2}+\frac{1}{z_0^2}\right)=\frac{\pi^2 D}{x_0^2}(1+\alpha^2+\beta^2).$$

Since the simple one-dimensional treatment gives $(1/\tau_0)-(1/\tau)=\pi^2D/x_0^2$ from (19), it is obvious that in this case

$$\bar{x}_0=\frac{x_0}{(1+\alpha^2+\beta^2)^{\frac{1}{2}}} \qquad (s\gg 2D/x_0). \tag{28}$$

To summarize the preceding arguments, if $s<2D/z_0$, a virtual thickness $\bar{x}_0=x_0/(1+\alpha+\beta)$ is used to replace x_0 in (20) in order to calculate s; if $s>2D/x_0$ a virtual thickness $\bar{x}_0=x_0/(1+\alpha^2+\beta^2)^{\frac{1}{2}}$ is used. Since in any practical experimental case $\alpha\ll 1$, $\beta\ll 1$, these corrections are not very large. If s lies between these two values, a correction intermediate in value between the two cases discussed above, arrived at by some manner of averaging, may be used. Note that for both corrections $\bar{x}_0\rightarrow x_0$ as α and β approach zero. Some slight foreknowledge of the order of magnitude of s is required to know which correction to apply. This is usually obvious from the relative magnitude of τ and the experimentally observed τ_0, and in any case can be calculated with sufficient accuracy to determine which correction is applicable by arbitrarily choosing one of the corrections (27) or (28) and substituting in (20). The values of $\bar{x}_0$ given by (27) and (28) are always close enough together to indicate unambiguously which correction should be applied.

To apply (20) to determine surface recombination velocity, in addition to the experimentally measured decay constant τ_0, it is necessary to know the bulk lifetime τ and the diffusion constant D for minority carriers. These two quantities may be determined experimentally by making minor changes in technique which are described below.

EXPERIMENTAL TECHNIQUES

A block diagram of the experimental equipment is shown in Fig. 1. A pulse generator produces a sharp pulse of 5 μsec duration and 300-volts amplitude across the germanium sample. Type 4C35 hydrogen thyratrons are utilized in the pulse generator to produce an output impedance of the order of 100 ohms. The pulse produced is of sufficient amplitude and duration to inject minority carriers and to sweep them uniformly throughout the sample. The pulse repetition rate is 60 sec^{-1}. The conductivity of the sample is measured subsequent to the pulse by measuring the voltage across a resistor in series with the germanium and a current source, and the measuring current is in most cases kept sufficiently small so that the proportion of carriers swept out of the sample by the conductivity measuring field to the total number of injected carriers is small. For samples of quite long lifetime, however, it is often necessary to apply a correction to account for carriers swept out of the sample in this manner. This correction will be discussed further below. The 1-V diode is inserted in the circuit to prevent the full pulse voltage from being applied to the input of the amplifier. The indicating instrument is a Dumont type 256-D oscilloscope which is equipped with triggered sweeps of various lengths and with calibrated time-base markers. The sweep is initiated by the injecting pulse. In practice, the time constants are arrived at by measuring photographs of the oscilloscope traces. It is felt that time constants can be measured in this manner with about ±5 percent accuracy.

Measurements of bulk lifetime are made using samples cut directly from zone-melted single crystal ingots and surface-treated to reduce surface recombination to an absolute minimum. The very large size of these samples (10×10×20 mm) and consequent small surface area-to-volume ratio, in conjunction with the low surface-recombination etch which is used, enable one to observe directly the bulk lifetime τ as the decay constant of the observed lifetime trace. Naturally, for samples of very long bulk lifetime it is not possible to neglect surface effects in this manner, but for the samples used in these investigations ($\tau\sim 200$ μsec) the discrepancy between the observed time constant and the actual bulk lifetime is well within the uncertainty of the measurements. This can be seen at once from (12).

Surface recombination velocities were determined for a large number of surface treatments by cutting small samples out of zone-melted single crystal ingots whose bulk lifetime had been measured by the method described above. The samples employed for this purpose were about 1 mm thick, 3 to 5 mm wide, and 10–15 mm long. The observed decay constant for injected carriers was obtained and s calculated using (20) with the appropriate virtual thickness $\bar{x}_0$ obtained from (27) or (28). In the case of n-type samples good injection usually cannot be obtained from the pressure-type contacts used in this experiment unless the contact ends of the sample are metal-plated. In these experiments copper-plating has been employed, but nickel-plating will do as well. Soldered contacts may be used but are more difficult to prepare and necessitate heating the sample, which may be undesirable. Care must, of course, be taken to plate only the ends of the sample, as contact of the plating solutions or plated film with the etched sample surfaces may produce inaccurate determinations of surface recombination velocity. If the plating is confined to the end of the sample, the effect of this treatment on the surface recombination velocity may be neglected, since carriers are swept away from one plated contact by the conductivity measuring field, and their disappearance at the other contact is corrected for, if necessary, by the normal correction for sweeping effect discussed below. Also, the plated area is small compared to the total surface area of the sample. In the case of p-type samples, however, electron injection is unsatisfactory from copper-plated contacts. Bare metal-to-semiconductor contacts are sufficient to effect satisfactory electron injection into p-type germanium. This type of contact on n-type samples, however, causes very unsteady or "noisy"

injection. It has been possible, in spite of these difficulties, to make some observations on samples with end contacts unplated, and then to plate the ends of the samples as described above and measure the decay constants again. It is observed that plating the end contacts does not measurably affect the observed time constants.

The preparation of a sample for surface recombination velocity measurements involves a number of operations. The sample is lapped with 800 grit alundum lapping compound on all six sides to remove all traces of previous surface treatment and incidental surface contamination. Then the surface is chemically etched and rinsed in water, nitric acid if the etch contains heavy metal ions, and methyl alcohol. The ends of the sample are then plated (n-type samples only) and the sample is ready for measurement.

In order to use (20) to calculate surface recombination velocity, it is necessary to know the value of D, the diffusion constant for minority carriers. This value is commonly assumed to be 92 cm^2 sec^{-1} for electrons and 44 cm^2 sec^{-1} for holes. However, since the diffusion constant is connected with the drift mobility μ through the Einstein relation,

$$D=\frac{\mu kT}{e_0}, \tag{29}$$

and since it is observed experimentally that drift mobility of minority carriers varies somewhat from one sample of germanium to another, it is desirable to measure D for the actual samples at hand. This can be done by lapping the surfaces of the sample and observing the decay constant exactly as for the surface recombination velocity measurement. For such surfaces, s is very large and the argument of the tangent function in (20) is very nearly equal to $\pi/2$. Setting this argument equal to $\pi/2$ and solving for D gives

$$D=\frac{\bar{x}_0^2}{\pi^2}\left(\frac{1}{\tau_0'}-\frac{1}{\tau}\right), \tag{30}$$

where τ_0' is the observed time constant for lapped surfaces, and $\bar{x}_0$ is of course given by (28). The results of some actual measurements carried out in this manner are tabulated in the following section. The Einstein relationship between drift mobility and diffusion constant has been experimentally verified recently at the Bell Telephone Laboratories.[3] It should be noted that the diffusion constant obtained from these measurements is the ambipolar diffusion constant. If the sample is in the near-intrinsic conductivity range the ambipolar diffusion constant may differ considerably from the diffusion constant for minority carriers in the extrinsic conductivity range.[4]

If the observed lifetime is quite long, or the applied sweeping voltage high, the effect of carriers swept out of the sample by the conductivity measuring field can no longer be ignored, and a correction to the observed lifetime must be made. To account rigorously for the effect of this field, a term $-\mu E_0\partial n/\partial z$ should be added to the left-hand side of (1) and the resulting equation solved using (2) and (12). Since the solution obtained without this term is itself very cumbersome, it is desirable to avoid the additional complication introduced by adding the field term, and instead derive a correction to the solution of (1) for the field term which is applicable for small values of the applied conductivity measuring field E_0, and which will account for the fact that carriers are removed from the sample at one of the end contacts. Since the sweeping field is applied along the z axis, the number of carriers left in the sample at time t will be proportional to $(z_0-vt)e^{-t/\tau_0}$, where v is the drift velocity of the carriers under the influence of the sweeping field. If this drift velocity is expressed in terms of the drift mobility and applied voltage V_0, and the factor (z_0-vt) approximated by an exponential, it can be shown that if $\mu V_0\tau_1/z_0^2\ll 1$, then

$$\frac{1}{\tau_0}=\frac{1}{\tau_1}-\frac{\mu V_0}{z_0^2}, \tag{31}$$

where τ_0 is the zero-field time constant and τ_1 is the time constant observed when the conductivity field has the value $E_0=V_0/z_0$.

EXPERIMENTAL RESULTS

Results of measurements of surface recombination velocity for a variety of surface treatments on n-type and p-type samples are given in Table I. The details of these surface treatments are listed below.

1. WAg (Westinghouse Silver Etch): 4 cc HF, 2 cc HNO_3, 4 cc H_2O containing 200 mg $AgNO_3$. Allow to stand for ½ hr before use. Normal etch time 1–2 min. Preferential to (111) surfaces. Gives rough surface with bright (111) reflections.

2. H_2O_2 (Superoxol etch): 1 cc HF, 1 cc H_2O_2 (30 percent), 4 cc H_2O. Allow to stand ½ hour before use. Preferential to (100) surfaces. Normal etching time 1–2 min. Gives fairly smooth but nonspecular surface.

3. CP4: 15 cc CH_3COOH, 25 cc HNO_3, 15 cc HF, 0.3 cc Br_2. Allow to stand ½ hr before use. Nonpreferential etch, gives bright specular surface. Normal etching time, 1 min.

4. Boiling water: Boil sample 5 min in distilled water. Small weight loss, dull surface.

5. Cu etch: 4 cc HF, 2 cc HNO_3, 4 cc H_2O containing 200 mg $Cu(NO_3)_2$. Allow to stand ½ hr before use. Nonpreferential etch, gives bright specular surface. Normal etching time 1–2 min.

6. CP4-A: Same as CP4 above but without Br_2.

7. Pd etch: 4 cc HF, 2 cc HNO_3, 1 cc $PdCl_2$ solution containing 5 mg $PdCl_2$. Allow to stand ½ hr before use.

[3] Transistor Teachers Summer School, BTL, Phys. Rev. **88**, 1368 (1952).

[4] W. van Roosbroeck, Phys. Rev. **91**, 282 (1953).

TABLE I. Surface recombination velocities (cm/sec).

Etch	n-type		p-type	
	s(range)	s(mean)	s(range)	s(mean)
WAg (2 min)	15– 77	31(+18,−11)	53– 330	160(+130,−70)
H_2O_2 (2 min)	18– 96	53(+33,−20)		
CP4 (1 min)	50– 220	100		
Boiling water		$>2\times10^4$		
Cu (2 min)	210–1000	450	800–5000	2800
CP4-A (2 min)	260–1240	490		
NaOH elec. (2 min)	120– 960	290		
Pd (1 min)	630– 780	660	80– 430	250
5:5:1 (1 min)	1600–2800	2100		
Ni elec. (2 min)	250– 670	380		
Chromic acid elec. (2 min)		$>2\times10^4$		
H_3PO_3 elec. (2 min)		$\sim10^4$		

Normal etching time, 1 min. Gives bright specular surface.

8. NaOH electrolytic: Ge sample anode, Ge cathode, 10 percent NaOH solution electrolyte. Current density, 160 ma/cm². Normal etching time, 1 min.

9. 5:5:1 etch: 5 cc HNO_3, 5 cc HF, 1 cc H_2O. Allow to stand ½ hr before use. Normal etching time, 1 min.

10. Ni electrolytic: Ge sample anode, Ni wire cathode, Ni plating solution ($NiCl_2 \cdot 6H_2O$, 300 g/l; H_3BO_3, 30 g/l; HCl to make pH=1) electrolyte. Current density, 230 ma/cm². Normal etching time, 2 min.

11. Chromic acid electrolytic: Ge sample anode, Ge cathode, electrolyte, 2.5 g CrO_3, 5 cc HNO_3, 20 cc H_2O. Allow to stand 5 minutes before use. Current density, 10 ma/cm². Normal etching time, 2 min.

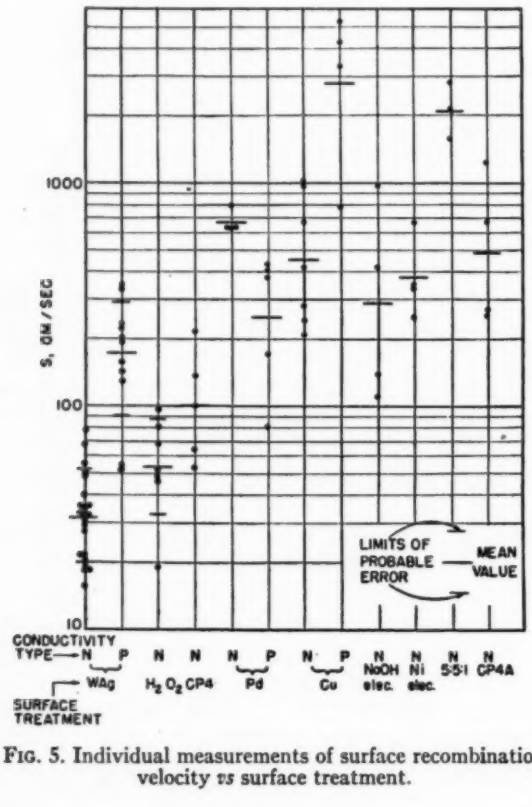


FIG. 5. Individual measurements of surface recombination velocity *vs* surface treatment.

12. H_3PO_3 electrolytic: Ge sample anode, Ge cathode, 10 percent H_3PO_3 electrolyte. Current density, 10 ma/cm². Normal etching time, 2 min.

The scatter of individual measurements of surface recombination velocity is shown in Fig. 5. For any given surface treatment there is usually a rather large spread of experimental surface recombination velocities measured with etching solutions of the same composition made up on different occasions and with different samples, sometimes as much as a factor of 2 above and below the mean value. This rather poor reproducibility is characteristic of surface phenomena of this type, since it is nearly impossible to reproduce exactly surface conditions from one trial to the next. These difficulties arise for several reasons. For instance, the etching agent requires an "induction period" to initiate its action, which varies from sample to sample. Etching action is usually initiated at small surface imperfections such as pits or scratches, and thus the microscopic condition of the surface, which it is impossible to reproduce exactly from one occasion to the next, plays an important rôle in the activity of the etching process.

Most of the etching solutions used in this investigation are dependent upon catalyzing agents such as heavy metal ions or halogen molecules for their activity; the actual mechanisms of the etching reactions themselves, however, are not well understood. For this reason it is impossible to say exactly what chemical conditions are necessary for reproducible etching, but it is observed that etch activity will vary from one batch of supposedly identical etching solution to another, thus adversely affecting reproducibility of surface conditions. In addition, small local variations in bulk lifetime from sample to sample may cause a scatter in calculated recombination velocities, which, however, should exhibit consistent errors on the same sample.

The action of some etches is preferential with respect to certain crystal planes. This leads one to expect that surface recombination velocity might vary as a function of crystal orientation for certain types of surface treatment. Experimentally, however, no correlation could be found between measured values of surface recom-

bination velocity and crystal orientation on n-type germanium samples oriented with (100), (110), and (111) planes in the large yz surface of the sample when etched with the (111) preferential WAg etch.[5]

It is observed that reproducibility is better for some surface treatments than for others. 22 measurements of surface recombination velocity using a 2-minute WAg etch on n-type samples gave a set of data whose rms deviation from the mean was less than ± 50 percent. This may be considered good reproducibility in this type of experiment. The WAg etch, in addition to providing very low surface recombination velocity, was felt to be superior to other low-recombination etches such as CP-4 and "Superoxol" in respect to reproducibility of surface conditions. In fact, one may utilize this good reproducibility observed for WAg etch to find bulk carrier lifetime directly from the decay constant τ_0 measured on small samples by assuming that $s=30$ cm/sec for n-type or 160 cm/sec for p-type samples and solving for τ in (23). If the difference between τ_0 and τ is fairly small compared to τ, one may obtain a fairly accurate estimate of bulk lifetime in this manner. One may assume that $D=90$ cm^2/sec for electrons and 45 cm^2/sec for holes in this calculation without introducing much additional error.

Note that in general surface recombination velocities are not the same for n-type and p-type samples using the same surface treatment. In the case of WAg etched surfaces, the values $s=30$ cm/sec for n-type and $s=160$ cm/sec for p-type material are in good agreement with the values obtained by Bardeen and Brattain[6] for a CP4 etched surface. On the other hand, in the case of the Palladium etch (No. 7 above) the values of s were lower for p-type samples than for n-type.

Values of the reflection probability R can be obtained from the experimental values of s by (13). This relation is shown graphically in Fig. 6 for $T=300°$K ($\bar{v}\sim 10^7$ cm/sec). $1-R$ represents the probability of recombination for collisions with the surface, and therefore can be construed as the effective surface trapping area per unit surface area. From this plot it is apparent that the surface recombination probability ranges from values of the order of 10^{-2} for poor surface treatments and lapped surfaces to about 10^{-5} for low surface recombination treatments such as WAg, CP4, etc. It is not unreasonable to suppose that the relatively large recombination probabilities associated with lapped surfaces may be explained by trapping of carriers and subsequent recombination in the lattice distortions and imperfections at the surface which are created by this treatment. It is found that when a lapped germanium surface is etched with WAg etchant, an amount of germanium corresponding to a layer 5μ thick ($\sim 10^4$ lattice constants in depth) must be removed to bring s down to the order of magnitude associated with the etched surface. These data provide some idea of the extent of the lattice disturbances mentioned above.

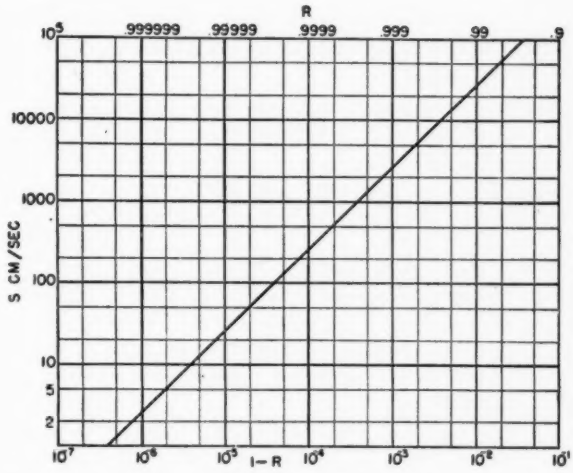


FIG. 6. Probability $1-R$ of recombination on collision with a surface *vs* surface recombination velocity.

Table II gives measured values of bulk lifetime, diffusion constant, and drift mobility of minority carriers for the ingots of germanium from which the samples used in the surface recombination measurements were cut. The values of drift mobility listed for these ingots are in reasonably good agreement with the values of Haynes and Shockley[7] which were measured by an entirely different method. For the samples listed in Table II the measured ambipolar diffusion constant is within about 2 percent of the actual extrinsic minority carrier diffusion constant.[4,*]

The authors wish to express their appreciation to Dr. J. W. Faust for his help in formulating surface treatments and etching the surfaces of the samples used in the experiments, and to Dr. C. W. Helstrom for helpful discussions concerning the diffusion problems encountered in the interpretation of the experimental data.

TABLE II. Drift mobility and bulk lifetime.

Ingot	Type	ρ, ohm-cm	D, cm^2/sec	μ cm^2/volt-sec	τ, μsec
Z- 24	n	11	53.4±6.8	2060±260	150±10
Z-103	n	9	40.0±2.6	1550±100	250±15
Z-105	n	11	45.9±5.7	1770±200	180±10
Z-100	p	20	78.6±6.1	3050±240	180±10

[5] R. H. Wynne and C. Goldberg, J. Metals **5**, 436 (1953).
[6] J. Bardeen and W. H. Brattain, Bell System Tech. J. **32**, 1 (1953).
[7] J. R. Haynes and W. Shockley, Phys. Rev. **81**, 835 (1951).
* *Note added in proof.*—Equation (18) represents strictly only the conductivity change due to injected minority carriers. Since simultaneous injection of majority carriers must take place in order to maintain over-all charge neutrality in the crystal, (18) should be multiplied by a factor $(b+1)$ to give the correct conductivity change, where b is the ratio of electron to hole drift mobility. Only the magnitude of (18) is affected by this factor, the time dependence, which is the only part of the expression used in the subsequent arguments, being unaffected.

Propagation of Microwaves between a Parallel Pair of Doubly Curved Conducting Surfaces*

K. S. KUNZ †
Electrical Engineering Department, Case Institute of Technology, Cleveland, Ohio
(Received May 25, 1953; revised manuscript received January 4, 1954)

Microwaves propagated between two doubly-curved parallel conducting surfaces, in the geometrical optics approximation, follow rays obeying Fermat's principle. The optical distance in this case is determined both by the index of refraction of the material between the surfaces and by the curvature of the surfaces. The focusing properties of such a parallel plate lens is fully specified by giving the metric of the two-dimensional optical space associated with the mean surface of the guide. A family of equivalent lenses is shown to exist, and formulas are obtained that permit one to go from any one of these lenses to any other. Furthermore, conditions are given for there to exist in a family of equivalent lenses a lens employing a unit index of refraction everywhere. Of particular interest is the demonstration that the Rinehart analog of the Luneberg lens and the RCA "R-2R" lens are members of a two-parameter family of lenses with analogous focusing properties.

1. INTRODUCTION

ONE of the techniques used in microwave antenna theory to achieve a certain desired focusing of an electromagnetic wave is to allow this wave to propagate between two closely-spaced conducting plates. Although other modes may be used, one usually makes the spacing between the plates small enough so that only the transverse electric and magnetic mode, the so-called TEM mode, is propagated. In this mode the velocity of propagation is independent of the spacing and equal to the velocity in free space, but one may achieve a focusing of the electromagnetic wave by using doubly-curved plates or joined sections of flat or singly-curved plates. The focusing depends on the relative path lengths along the various rays. This technique has been used chiefly in the design of antenna feeds such as those mentioned in references 7 and 9 (see Bibliography at the end of the paper).

The purpose of this paper is to develop the mathematical theory pertaining to the design of two-dimensional path-length lenses. This includes lenses in which the space between the plates is filled with a substance whose index of refraction varies from point to point. Disregarding the limitations imposed by practical considerations, we shall assume that the index of refraction n may be given by any positive function of the coordinates. The treatment will be confined to a geometrical optics approximation in which one determines the paths of the rays on a *mean surface*. It is clear that such a two-dimensional treatment demands in general that the spacing between the conducting surfaces be small compared to a wavelength and to the principal radii of curvature of the conducting surfaces; however, no attempt is here made to determine the range of applicability of this geometrical optics treatment.

This paper is related to the work of Myers, Levy, and Rinehart as given in the set of references at the end and represents an extension and generalization of their results.

2. TEM MODE OF PROPAGATION IN TWO-DIMENSIONAL GUIDES

As stated before, the conducting surfaces forming the guide are assumed to be parallel, i.e., they have common normals at all points.‡ Moreover, for theoretical purposes we assume that the spacing is so small that there is little difference between the curvatures of the two conducting surfaces, and one may reduce the problem to one of two dimensions. One defines a *mean surface*, which is the locus of the midpoints of the common normals, and treats the propagation within the guide in terms of rays drawn on this *mean surface*.

Usually we shall describe a guide only in terms of the shape of its *mean surface* and the index of refraction n, the latter being expressed as a function of curvilinear coordinates on this surface. To obtain an actual physical guide from this mathematical surface, one needs to construct the two physical surfaces. These surfaces are the locus of points lying a fixed distance from the mean surface. For example, in the Rinehart-Luneberg guide described later and designed for 3.2-centimeter operation this distance from the mean surface was taken as one-quarter of an inch.

3. COORDINATE SYSTEM FOR A SURFACE OF REVOLUTION

For the present we shall assume that all guides considered, as well as their indexes of refraction, are axially symmetric. Their mean surfaces are then surfaces of revolution, and n is a function only of the distance from the axis. Thus the mean surface is generated by rotating a curve C, called the generating curve, about some line

* Through sponsorship of the Electronics Research Division, Air Force Cambridge Research Center, Cambridge, Massachusetts.
† Now at the Research Laboratory, Schlumberger Well Surveying Corporation, Ridgefield, Connecticut.

‡ This requirement is made merely for the sake of preciseness, no harm would apparently be done by using a slightly different relationship between the surfaces. For instance, one might prefer to take for the conducting surface two neighboring surfaces of a one-parameter family of surfaces.

which we may take as the z axis of a cylindrical coordinate system, see Fig. 1. The other two cylindrical coordinates are ρ and θ.

Instead of describing C and hence the surface S in terms of the function $z(\rho)$, it is convenient to use the functional relationship between ρ and s, where s is the distance from the axis, or other suitable point, measured along C. Thus one may specify the surface by giving $\rho(s)$ or by giving $s(\rho)$. An important advantage of using the function $\rho(s)$ is that ρ is always a single-valued function of s. In contrast, $s(\rho)$ may readily be multiple valued. Even so, for many problems it is easier to use $s(\rho)$.

One may use as the surface coordinates either s and θ or, on occasions, ρ and θ. In any case the metric of the surface, which gives the distance dL between neighboring points, is expressed by

$$dL^2=ds^2+\rho^2d\theta^2. \tag{3.1}$$

If one multiplies this distance dL by n, the index of refraction, one obtains, by definition, the *optical distance* $d\zeta$ between neighboring points (for the TEM mode of propagation); thus

$$d\zeta^2=n^2[ds^2+\rho^2d\theta^2]. \tag{3.2}$$

If s and θ are taken as the independent variables, Eq. (3.2) may be written out more explicitly as

$$d\zeta^2=n(s)^2[ds^2+\rho(s)^2d\theta^2], \tag{3.3}$$

while if ρ and θ are used, it may be written

$$d\zeta^2=N(\rho)^2[s'(\rho)^2d\rho^2+\rho^2d\theta^2], \tag{3.4}$$

where $N(\rho)=n(s)$.

4. FERMAT'S PRINCIPLE GENERALIZED

The whole of geometrical optics may be summarized by Fermat's principle which states that the actual path of a ray between two points A and B is such as to make the variation of the optical path length zero. Usually this means that the ray takes the path of minimum optical length. This principle accounts for the laws of reflection and refraction in space. It may also, of course, be specialized to two-dimensional optical problems described by rays in a plane.

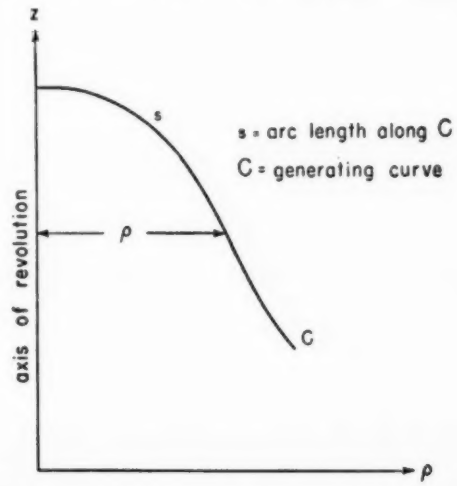


FIG. 1. Generating curve of the mean surface of a guide.

We shall generalize the principle to the extent of assuming that it holds for the rays on the curved surface representing the mean surface of a two-dimensional guide. Thus the path of a ray between two points A and B on the mean surface S is such as to make

$$\delta\int_A^B d\zeta=\delta\int_A^B ndL=0. \tag{4.1}$$

If the index is constant everywhere,

$$\delta\int_A^B dL=0, \tag{4.2}$$

and thus the rays are geodesics on the surface.§

5. EQUIVALENT GUIDES

If the optical metrics [see Eq. (3.2)] for two different guides may be set equal by assuming the proper functional relation

$$s'=f(s,\theta),\quad \theta'=g(s,\theta) \tag{5.1}$$

between the coordinates on their mean surfaces, one may refer to the two as *equivalent guides*. These functional relations between the coordinates (s,θ) and (s',θ') provide one with a point P' on the mean surface S' of one guide corresponding to each point P on the mean surface S of the other. One has thus a point-by-point mapping of the S on S'. This mapping preserves the optical distances between neighboring points as given by the metric in (3.2), and hence the line integrals of the optical distance along equivalent paths, i.e., along paths joining equivalent points, are equal.

Since optical distances along corresponding paths on S and S' are equal, it is clear by Eq. (4.1) that the rays describing the propagation of the electromagnetic field will pass through equivalent points, the paths being in each case extremal (usually minimum) optical paths. That is, the mapping functions (5.1) may be applied directly to the rays. The rays will thus exhibit the same topological relationship to one another in equivalent guides. One may summarize these statements by saying that the optical goemetry is the same for all equivalent guides.

6. EQUIVALENCE FOR AXIALLY SYMMETRICAL GUIDES

Since we are considering for the present only guides having complete axial symmetry, the optical metrics are of the form of Eq. (3.3), or Eq. (3.4). If we use bars over quantities to indicate those relating to a particular

§ For experimental confirmation of this theory, see references 2 and 3.

guide and unbarred quantities for those of any other equivalent guide, then the equivalence of the optical metrics, by Eq. (3.2), requires that

$$n^2(ds^2+\rho^2d\theta^2)=\bar{n}^2(d\bar{s}^2+\bar{\rho}^2d\bar{\theta}^2). \tag{6.1}$$

One may achieve this result by requiring that

$$nds=\bar{n}d\bar{s}, \quad n\rho=\bar{n}\bar{\rho}, \quad \theta=\bar{\theta}. \tag{6.2}$$

This leads to the following set of integral relationships,

$$\int_a^s nds=\int_{\bar{a}}^{\bar{s}}\bar{n}d\bar{s}\equiv A(\bar{s}) \tag{6.3a}$$

$$n\rho=\bar{n}\bar{\rho}\equiv B(\bar{s}) \tag{6.3b}$$

$$\int_b^s ds/\rho=\int_{\bar{b}}^{\bar{s}}d\bar{s}/\bar{\rho}\equiv C(\bar{s}), \tag{6.3c}$$

where a and $\bar{a}$, b, and $\bar{b}$ represent values of s and $\bar{s}$ corresponding to equivalent points in the two guides.

Now $\bar{n}(\bar{s})$ and $\bar{\rho}(\bar{s})$ are known, since they correspond to the given guide. If one chooses any other mean surface described by $\rho(s)$, one may, in principle at least, determine the index of refraction $n(s)$ needed to make the guides equivalent as follows. One obtains the mapping function $\bar{s}=f(s)$ by solving Eq. (6.3c) for $\bar{s}$. From Equation (6.3b), therefore,

$$n(s)=B(f(s))/\rho(s). \tag{6.4}$$

On the other hand, if $n(s)$ is specified, one obtains $f(s)$ from the first equation and then $\rho(s)$ from Eq. (6.3b). Since for any actual surface, see Sec. 12, $|d\rho/ds|\leqslant 1$, the $\rho(s)$ obtained above need not represent a realizable surface.

7. FAMILY OF GUIDES EQUIVALENT TO THE LUNEBERG LENS

A flat guide with index of refraction given in terms of the radius $r=\bar{s}$ by

$$\bar{n}(r)=\begin{cases}(2-r^2)^{\frac{1}{2}}, & \text{for } r\leqslant 1\\ 1 \quad\quad , & \text{for } r\geqslant 1\end{cases} \tag{7.1}$$

has the property that rays from a source on the unit circle $r=1$ will be focused into a parallel beam in the region $r\geqslant 1$. For practical considerations one eliminates the portion of the guide outside the unit circle letting the electromagnetic field spread normal to the plane. This does not seriously influence the beam in the plane of the surface. The flat guide within the unit circle with index $(2-r^2)^{\frac{1}{2}}$ is referred to as the *Luneberg lens*.[11] The restriction of radius to unit length is no real restriction since one may use any arbitrary length as the unit of length.

If one finds a curved plate guide whose outer edge is a circle of unit radius that has the same optical geometry as the Luneberg lens, one may substitute this new guide for the Luneberg lens in the complete guide of (7.1). What is more, the new guide will exhibit the same focusing property as the Luneberg lens when the outer guide is omitted and one operates the lenses in free space.

One may obtain such an equivalent guide corresponding to any arbitrary shape described by $\rho(s)$ by determining the proper $n(s)$. Since for a flat guide $\bar{s}=\bar{\rho}=r$, Eq. (6.3c) becomes

$$\int_b^s ds/\rho=\int_1^r dr/r=\ln r. \tag{7.2}$$

Thus the mapping function is

$$r=f(s)=\exp\int_b^s ds/\rho, \tag{7.3}$$

where $s=b$ corresponds to $r=1$. Since by Eqs. (6.3b) and (7.1), $B(\bar{s})=r(2-r^2)^{\frac{1}{2}}$, one has by Eq. (6.4)

$$n(s)=f(s)[2-f(s)^2]^{\frac{1}{2}}/\rho(s). \tag{7.4}$$

If one employs ρ as the independent variable in place of s, then

$$F(\rho)=\exp\int_1^\rho[s'(\rho)/\rho]d\rho \tag{7.5}$$

replaces $f(s)$ in Eq. (7.4), and $N(\rho)$ is written in place of $n(s)$.

On the other hand, if one specifies $N(\rho)$, the required shape for equivalence to the Luneberg lens is obtained as follows. Solving Eq. (6.3b), which is now

$$\rho N(\rho)=r(2-r^2)^{\frac{1}{2}}, \tag{7.6}$$

for r, keeping in mind that $0\leq r\leq 1$, one obtains the mapping function

$$r=F(\rho)=[1-(1-\rho^2N^2)^{\frac{1}{2}}]^{\frac{1}{2}}. \tag{7.7}$$

Next differentiating Eq. (7.1) with respect to ρ, one finds that

$$s'(\rho)/\rho=(dr/d\rho)r^{-1}. \tag{7.8}$$

From Eq. (7.7) therefore

$$s'(\rho)=\tfrac{1}{2}[1+(1-\rho^2N^2)^{-\frac{1}{2}}][1+\rho N'/N]. \tag{7.9}$$

The integral of this is the required shape function $s(\rho)$.

Example 1.—Determine the surface having $N(\rho)=1$ which is equivalent to the Luneberg lens.

From Eq. (7.9)

$$s(\rho)=\tfrac{1}{2}\int_0^\rho[1+(1-\rho^2)^{-\frac{1}{2}}]d\rho=\tfrac{1}{2}(\rho+\sin^{-1}\rho), \tag{7.10}$$

the constant of integration being determined such that

[11] This is the two-dimensional equivalent of the perfect optical lens discussed by R. K. Luneberg, see reference 6.

TABLE 1. Lenses theoretically equivalent to the Luneberg lens.

Surface	Generating curve	Index needed $[n(s)$ or $N(\rho)]$	Permissible range
1. Plane (Luneberg lens itself)	$s=\rho$	$[2-\rho^2]^{\frac{1}{2}}$	$0\leq\rho\leq 1$
2. Cone	$s=k\rho$	$\rho^{k-1}[2-\rho^{2k}]^{\frac{1}{2}}$	$0\leq\rho\leq 1$ if $k>0$ $\rho\geq 1$ if $k<0$
3. Cylinder	$\rho=1$	$e^{-s}[2-e^{-2s}]^{\frac{1}{2}}$	$s\geq 0$
4. Rinehart surface	$s=\frac{1}{2}(\rho+\sin^{-1}\rho)$	1	$0\leq\rho\leq 1$
5. Generalized Rinehart surface	$s=(1-k)\rho+k\sin^{-1}\rho$	$\dfrac{\{2[1+(1-\rho^2)^{\frac{1}{2}}]^{2k}-\rho^2\}^{\frac{1}{2}}}{[1+(1-\rho^2)^{\frac{1}{2}}]^{2k}}$	$0\leq\rho\leq 1$
6. Hemisphere	$s=\sin^{-1}\rho$	$\dfrac{\{2[1+(1-\rho^2)^{\frac{1}{2}}]^{2}-\rho^2\}^{\frac{1}{2}}}{[1+(1-\rho^2)^{\frac{1}{2}}]^{2}}$	$0\leq\rho\leq 1$

$s=0$ on the axis. This is the mean surface of the Rinehart analog of the Luneberg lens.¶

By Eq. (7.7) the mapping function giving the correspondence between points in the Luneberg lens and the Rinehart analog is

$$r=[1-(1-\rho^2)^{\frac{1}{2}}]^{\frac{1}{2}}. \tag{7.11}$$

Thus, in particular, $r=0$ for $\rho=0$, and $r=1$ for $\rho=1$.

8. RELATIONSHIP BETWEEN $N(\rho)$ AND $s(\rho)$ AT THE EDGE

A rather striking relationship exists for the above family of guides between the index of refraction $N(\rho)$ and the slope of the generating curve, as described by $s(\rho)$, at $\rho=1$.

From Eq. (7.8)

$$s'(\rho)=\frac{1+(1-\rho^2N^2)^{\frac{1}{2}}}{2N}\,\frac{N+\rho N'}{(1-\rho^2N^2)^{\frac{1}{2}}},$$

and since by Eq. (7.5) (recalling that $r=1$ when $\rho=1$)

$$N(1)=1, \tag{8.1}$$

one has

$$s'(1)=\tfrac{1}{2}K, \tag{8.2}$$

where

$$K=\lim_{\rho\to 1}\frac{N+\rho N'}{(1-\rho^2N^2)^{\frac{1}{2}}}. \tag{8.3}$$

Since the denominator in Eq. (8.3) vanishes, K and hence $s'(1)$ is infinite unless the numerator also vanishes. The numerator vanishes only if $N'(1)=-1$. Thus $N'(1)=-1$ unless the surface has a vertical tangent at the edge.

If $N'(1)=-1$, one evaluates K by L'Hospital's rule. Thus

$$K=\lim_{\rho\to 1}\frac{2N'+\rho N''}{-\rho N\left[\dfrac{N+\rho N'}{(1-\rho^2N^2)^{\frac{1}{2}}}\right]}=\frac{2-N''(1)}{K},$$

and therefore by Eq. (8.2)

$$N''(1)=2-K^2=2[1-2s'^2(1)]. \tag{8.4}$$

A surface having a horizontal tangent at the edge will have $s'(1)=1$ and hence $N''(1)=-2$. In fact for such a surface

$$N(1)=1,\quad N'(1)=-1,\quad N''(1)=-2. \tag{8.5}$$

9. CONICAL SKIRT TERMINATION

As seen in Sec. 7, the Rinehart-Luneberg (RL) lens, which is the geodesic (or path length) analog of the Luneberg variable dielectric lens, has a vertical tangent at its edge. This means that the electromagnetic wave on leaving the lens would be propagated vertically. To obtain horizontal propagation, as required by the optics, a toroidal bend may be used, see Fig. 2. Since (see Sec. 7) the transition from the mean surface of the RL lens to the horizontal plane representing the propagation in free space should be abrupt, the toroidal bend introduces an optical error and the rays do not emerge exactly parallel.

One may minimize the optical effect of the toroidal bend by making its radius small and by refocusing; i.e., by moving the source up into the lens to compensate in part for the converging effect of the bend. On the other hand one may, within the limits of the geometrical optics that considers the propagation to be along rays on the mean surface, achieve perfect optics by replacing the toroidal bend by the conical skirt termination shown in Fig. 2. In this device the rays on the mean

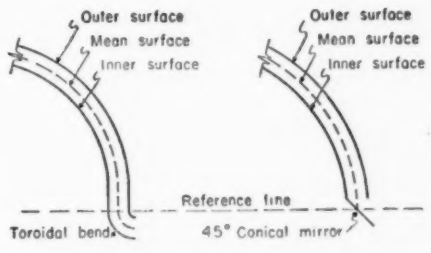


FIG. 2. Generating curves for the toroidal bend and conical skirt termination of the Rinehart-Luneberg lens.

¶ See references 9 and 10.

surfaces are turned abruptly into the horizontal plane by a 45° conical mirror formed by the inner conducting surface.

From the point of view of electromagnetic theory the conical skirt, if indefinitely extended, forms an image of the aperture. This image will obviously have the electric and magnetic vector in proper directions for propagation in a horizontal plane; moreover, the fields of the aperture and its image will be at right angles so that one may expect only a very small amount of coupling between them. (This also means that one would not expect the termination, in itself, to cause serious reflections.)

Tests on a lens with conical skirt termination indicate that this termination works somewhat better than the toroidal bend.

10. GEODESIC ANALOG OF A FLAT GUIDE WITH VARIABLE INDEX

In this section we shall obtain convenient expressions for a unit-index guide that is optically equivalent to a given flat guide with a variable index of refraction $n(r)$. From Eqs. (6.3), since $\bar{\rho}=\bar{s}=r$ and $n(s)=1$,

$$s=\int_{\bar{a}}^{r} n(r)dr \tag{10.1}$$

$$\rho=rn(r) \tag{10.2}$$

$$\int_{b}^{s} ds/\rho=\int_{1}^{r} dr/r=\ln r. \tag{10.3}$$

By this choice of lower limits, $s=0$ is made to correspond to $r=\bar{a}$ and $s=b$ to $r=1$. Solving Eq. (10.3) for r,

$$r=\exp\int_{b}^{s} ds/\rho=\exp\int_{c}^{\rho} s'(\rho)d\rho/\rho. \tag{10.4}$$

If $n(r)$ is known for the flat guide, Eqs. (10.1) and (10.2) express in parametric form the relationship between the s and ρ of the curved guide. One may, of course, obtain an explicit expression for $s(\rho)$ or $\rho(s)$ by eliminating the parameter r from between these equations.

On the other hand, if $\rho(s)$ or $s(\rho)$ is given, one obtains immediately from Eq. (10.4) the mapping function $r=f(s)$ or $r=F(\rho)$, and thus from Eq. (10.2)

$$n(r)=\rho/r=F^{-1}(r)/r=\rho(f^{-1}(r))/r. \tag{10.5}$$

Example 2.—Determine the geodesic analog of a plane with index $n(r)=r^{-k}$. By Eqs. (10.1) and (10.2), using the lower limit $\bar{a}=1$,

$$s=\int_{1}^{r} r^{-k}dr=(r^{1-k}-1)/(1-k) \tag{10.6}$$

and

$$\rho=r^{1-k}; \tag{10.7}$$

therefore,

$$s=(\rho-1)/(1-k). \tag{10.8}$$

This represents a cone with a slant angle ψ, the angle an element makes with the horizontal plane, given by

$$\sec\psi=s/(\rho-1)=(1-k)^{-1}. \tag{10.9}$$

In order for Eq. (10.8) to represent a realizable surface one must have $|\sec\psi|\geqslant 1$; therefore, $|1-k|\leqslant 1$ or $0\leqslant k\leqslant 2$.

Example 3.—Find what $n(r)$ is required in a plane to produce a complete analog of a spherical surface with unit index.

For a sphere of unit radius

$$\rho=\sin s, \tag{10.10}$$

and therefore by Eq. (10.4)

$$r=\exp\int_{b}^{s}\csc s ds=a\tan(s/2), \tag{10.11}$$

where $a=\cot(b/2)$. Thus, $s=0$, $s=\frac{1}{2}\pi$, and $s=\pi$ correspond respectively to $r=0$, $r=a$, $r=\infty$. By Eq. (10.6)

$$n(r)=r^{-1}\sin[2\tan^{-1}(r/a)]=2a(a^2+r^2)^{-1}. \tag{10.12}$$

The arbitrary parameter a can be given any positive value one chooses. Thus, there are an infinite number of flat surfaces having the same optical geometry. They are, however, very simply related. Since from Eq. (10.12)

$$an(r)=2[1+(r/a)^2]^{-1}, \tag{10.13}$$

one sees that $an(r)$ is always related in the same way to r/a. Thus one can scale down any particular solution by dividing by a if at the same time one multiplies the index of refraction by a. We shall refer to this transformation as a *transformation of scale.*

All rays starting from some point Q on the unit sphere having coordinates (s,θ) will pass through the image point P at the extremity of the same diameter and having coordinates $(\pi-s,\theta+\pi)$. Likewise by virtue of the mapping in Eq. (10.11) all rays from a point Q' in the plane with coordinates (r,θ) will pass through the image point P' having coordinates $(a^2/r,\theta+\pi)$.

The flat plane with index $n(r)=2a/(a^2+r^2)$ is *Maxwell's fisheye.*

11. SEMIPHYSICAL PRESENTATION OF THE EQUIVALENCE BETWEEN GUIDES

Consider a flat guide with index $n(r)$ and its geodesic analog. All points in a ring of the first lying between r and $r+\Delta r$ will be mapped on a ring of the second of width Δs in such a way that all optical distances are preserved. This means, in particular, that the optical thickness of the rings and their optical circumferences are the same. Mathematically, therefore,

$$\begin{aligned}\Delta s&=n(r)\Delta r\\ 2\pi\rho&=2\pi rn(r),\end{aligned} \tag{11.1}$$

which are seen to be essentially Eq. (10.2) and the differential form of Eq. (10.1). Finally Eq. (10.3) may be obtained by dividing the first of these equations by the second. Optical equivalence is, therefore, achieved by preserving the optical thickness and circumference of corresponding rings.

One can thus visualize the mean surface of the Rinehart-Luneberg lens as that surface produced by taking the rings of the flat Luneberg surface relaxing their associated index of refraction to unity and at the same time letting them expand so as to keep their optical thickness and circumference unchanged. Note that the outermost ring, since it has an index of unity, will not expand. It will, however, stand on edge since its inner and outer edges have, neglecting higher orders of the differentials, the same circumference.

It is clear from these considerations that the optical geometry of a guide is fully defined by the functional relation between the optical circumference and the optical distance from the axis, or some reference circle, along the surface. Letting $2\pi\sigma$ represent the former and τ the latter, the optical properties of the guide are thus fully defined by the function $\sigma=\sigma(\tau)$, and in particular

$$\sigma=n(s)\rho(s), \quad \tau=\int_0^s n(s)ds. \tag{11.2}$$

12. REALIZABILITY OF A UNIT-INDEX GUIDE

It is of interest to inquire if an equivalent unit-index guide, or geodesic analog, exists for every flat guide employing a variable index $n(r)$ and such a flat guide for every unit-index guide.

For a guide with unit index of refraction, Eqs. (11.2) become

$$\sigma=\rho \quad \text{and} \quad \tau=s, \tag{12.1}$$

and hence $\sigma(\tau)$ becomes simply $\rho(s)$. From Fig. 3 we have $d\rho=ds\cos\gamma$, and thus $|d\rho/ds|\leqslant 1$. Or in terms of $\sigma(\tau)$

$$-1\leqslant d\sigma/d\tau\leq 1. \tag{12.2}$$

Consider next a flat surface with an index $n(r)$ that is any arbitrary positive function of r. For such a space

$$\sigma=rn(r), \quad \tau=\int_0^r n(r)dr; \tag{12.3}$$

therefore,

$$d\sigma/d\tau=[n(r)+rn'(r)]/n(r) \tag{12.4}$$

or

$$n'(r)=[d\sigma/d\tau-1]n(r)/r. \tag{12.5}$$

Since $n'(r)$ can have any value from $-\infty$ to ∞, no restriction is imposed on $d\sigma/d\tau$ in such a guide.

If a flat guide is to have a unit-index guide as an analog, $d\sigma/d\tau$ satisfies Eq. (12.2). Therefore, by Eq. (12.5), since $n(r)$ is positive,

$$-2/r\leqslant n'(r)/n(r)\leqslant 0. \tag{12.6}$$

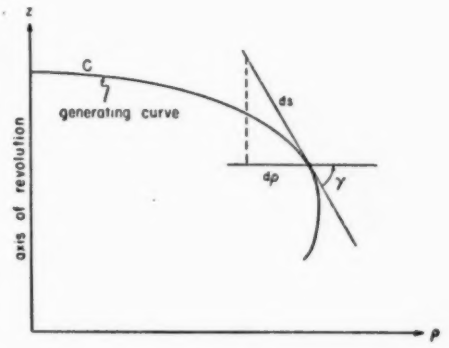


FIG. 3.

Integrating this inequality between any arbitrary r_0 and $r>r_0$, one obtains the inequality $2\ln(r_0/r)\leqslant\ln[n(r)/n(r_0)]\leq 0$ or its equivalent

$$n(r_0)r_0^2/r^2\leq n(r)\leq n(r_0). \tag{12.7}$$

Thus if one is to find a unit-index guide whose mean surface is a surface of revolution** that is equivalent to a flat guide with index of refraction $n(r)$, this $n(r)$ must not increase with r or decrease faster than $1/r^2$.

The above limitation can be understood in the following way. Take the limiting case of $n(r)$ being constant at $r=r_0$ and consider the infinitesimal ring (see Sec. 11) at this distance. Since the index is constant throughout this ring, it behaves as it would if it were a part of a plane with constant index of refraction, the geodesic analog of which is again a plane. The ring thus expands uniformly in proportion to the values of $n(r_0)$.

If $n(r)$ were to increase with r at $r=r_0$, upon relaxing the index to unity the outside edge of the ring would need to be expanded proportionally more than the inside to preserve the optical distances, this would cause the outer edge to become ruffled. The ring could not therefore be a part of a surface of revolution with unit index.

The other limiting case if for $n(r)$ to decrease as $1/r^2$ at $r=r_0$. The ring at $r=r_0$ then behaves as part of a plane with index given by c^2/r^2, c being an appropriate constant. Since the analog of such a plane is the same plane inverted in the circle $r=c$, see Example 2 for $k=2$, the relaxation of the index to unity will cause the above ring to reverse its inner and outer edges. Any faster decrease in the index will reduce the optical circumference of outside edge of ring still further relative to inside edge. The relaxation of the index will then not only reverse inner and outer edges but will cause a ruffled, or scalloped, edge to be formed, which again is not permissible for a surface of revolution.

13. NONISOTROPIC ANALOG OF THE RINEHART-LUNEBERG LENS

The functional relationship between s and ρ does not define a unique surface since a reflection of a part of the

** The statement is not true if one permits the unit-index guide to have a more general type of mean surface.

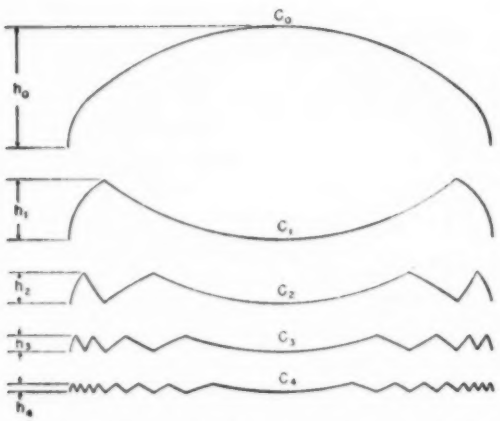


FIG. 4. A family of curves having the same functional relationship between s and ρ, namely: $s=\frac{1}{2}(\rho+\sin^{-1}\rho)$.

surface in a horizontal plane intersecting the surface does not change this relationship. In Fig. 4 is shown a family of curves C_0, C_1, C_2, etc., all of which have the same relation between arc length s and radial distance ρ, namely,

$$s=\tfrac{1}{2}(\rho+\sin^{-1}\rho). \tag{13.1}$$

The curve C_0 is the generating curve of the mean surface of the Rinehart-Luneberg lens; its height is lettered h_0. If one draws a horizontal line through C_0 at a height $h_0/2$ and reflects the portion of C_0 above it in this line, one obtains C_1. Similarly by reflecting the curve C_1 in a line drawn at half its height h_1, one obtains C_2. The process can be continued indefinitely. Each of the curves can be used to generate a surface of revolution that can theoretically be used as a mean surface of a lens.

The height of the curve C_n is $h_n=h_0/2n$, which approaches zero as n approaches infinity. The end result is a generating curve lying along a straight line but whose differential length between a point A at ρ and a point B at $\rho+d\rho$ is $s'(\rho)d\rho$. The significance of the mean surface in an optical sense for such unit-index guides is that it determines the optical distance between points, this distance being measured along geodesics. Therefore, the surfaces generated by the limiting curve of the family is a flat sheet on which the metric is

$$d\zeta^2=s'^2(\rho)d\rho^2+\rho^2d\theta^2, \tag{13.2}$$

the same as for the Rinehart-Luneberg lens—that generated by C_0. One may, however, interpret this metric as the optical distance between neighboring points in a nonisotropic dielectric slab with index of refraction

$$N(\rho)=\begin{cases}s'(\rho), \text{ for radial rays}\\ 1, \text{ for rays perpendicular to the radius.}\end{cases} \tag{13.3}$$

The argument is only used to suggest the nature of the relationship. The equivalence of the metric for the nonisotropic slab and that for the Rinehart-Luneberg lens insures their optical equivalence.

Practically, one might construct an approximation to such a nonisotropic lens by using obstructions in the form of circular ridges as suggested by the curves C_n. Any means, however, may be used that slows down the rays only in the radial direction. One serious difficulty, however, is the singularity of $s'(\rho)$ at $\rho=1$, see Eq. (13.1).

14. GEODESICS ON SURFACES OF REVOLUTION

From Eqs. (3.1) and (4.2) the equation for a geodesic on a surface of revolution is given in variational form by

$$\delta\int dL=\delta\int[ds^2+\rho^2d\theta^2]^{\frac{1}{2}}$$
$$=\delta\int[s'^2(\rho)+\rho^2\theta'^2(\rho)]^{\frac{1}{2}}d\rho=0, \tag{14.1}$$

where the primes indicate derivatives with respect to ρ. Applying Euler's equation (from the calculus of variations) to the dependent variable θ and its derivative θ'.

$$\rho^2d\theta/dL=h=\text{constant}, \tag{14.2}$$

or in terms of $d\theta$ and ds

$$d\theta=hds/[\rho(\rho^2-h^2)^{\frac{1}{2}}]. \tag{14.3}$$

From Fig. 5 one sees that

$$\sin\alpha=\rho d\theta/dL, \tag{14.4}$$

where α is the angle between the geodesic and the meridian; therefore Eq. (14.2) may be written

$$\rho\sin\alpha=h. \tag{14.5}$$

This is the fundamental equation for the behavior of a geodesic on a surface of revolution. For certain purposes, however, one may use the alternative form given in Eq. (14.3).

At the point of closest approach of the geodesic to the axis $\alpha=\pi/2$; therefore h is equal to the value of ρ at that point.

If one imagines a particle of unit linear momentum moving along the geodesic, $\rho\sin\alpha$ is its angular momentum. In terms of such a particle the fundamental Eq. (14.5) is just a statement of the conservation of angular

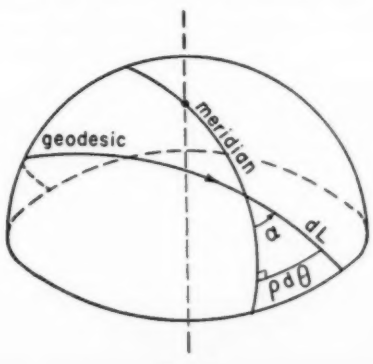


FIG. 5.

momentum. One can, in fact, show that a mass particle, constrained to move on the surface of revolution, but otherwise free, will move along geodesics on the surface. This follows from the fact that the force acting on the particle is normal to the surface and thus furnishes no torque to change the angular momentum of the particle.

15. A GENERAL FAMILY OF LENSES††

Consider a surface of revolution S (see Fig. 6) generated by a curve C whose arc length s is given by

$$s=A\rho+B\sin^{-1}\rho, \tag{15.1}$$

where A and B are constants. Let this surface S be attached to a cone K along the unit circle $\rho=1$. It will now be shown that all the geodesic rays passing on the same side‡‡ of axis of revolution can be made to be parallel in a plane representing the development of the cone K if the following conditions are met:

(1) The source of the geodesic rays is any point P on the unit circle $\rho=1$, hereafter referred to as the *feed circle.*

(2) The slant angle ψ of the cone is taken to be

$$\psi=\sec^{-1}2A. \tag{15.2}$$

This limits the choice of A, since clearly $|A|$ must be greater than $\frac{1}{2}$.

(3) The rays considered must pass through S.

One may choose as an aperture any convenient straight line in the development of K, and for a proper choice of this aperture one can presumably deform the cone in the neighborhood of this aperture so that the aperture becomes a straight line. This is the same idea used in the RCA "R-2R" Lens (see reference 1). The guide is terminated at this aperture and from there on the electromagnetic wave is propagated in free space.

If a ray starts from the point P in a direction making an angle α_1 with the meridian through P, then on passing

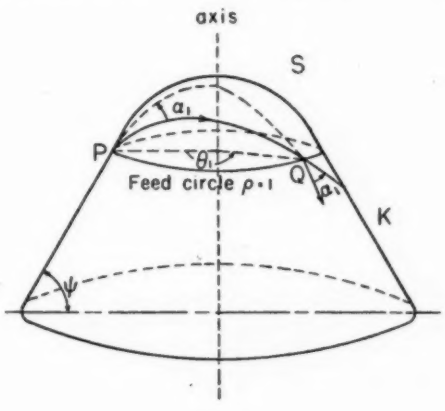


FIG. 6.

†† Professor Wade Ellis of Oberlin College and Mrs. E. C. Fine, who was then at the Air Force Cambridge Research Center, worked closely with the author in obtaining the results of this section.

‡‡ The rays passing to the right and to the left of the axis may form two separate beams.

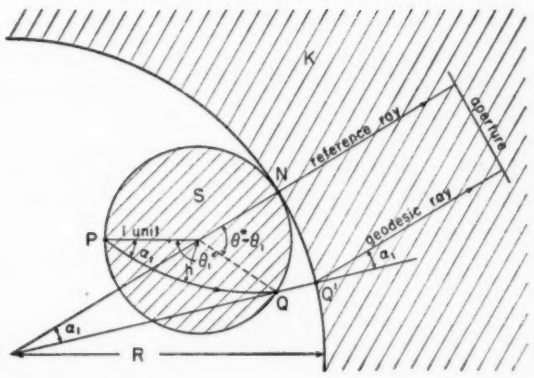


FIG. 7. Path of a general geodesic ray when the cone K is developed.

through S it will emerge from S at some point Q where, by Eq. (14.5), it will make the same angle α_1 with meridian through Q. If one designates by θ_1 the number of degrees in the arc PQ of the feed circle measured in the direction the ray circles the axis§§ then the shape of the surface S will determine the functional relation between θ_1 and α_1.

In order for the rays to be parallel in the development of the cone K, see Fig. 7, one must have

$$\theta^*-\theta_1=\alpha_1 R, \tag{15.3}$$

where R is the radius of circular arc NQ' and θ^* is the value of θ, for the limiting ray corresponding to $\alpha_1\to 0$. This insures the equality of arc lengths NQ and NQ', thus permitting the joining of the pair of points Q and Q' in forming the cone.

Since, as shown earlier, h is the distance of the geodesic ray from the axis at its nearest point, by Eq. (14.3)

$$\tfrac{1}{2}\theta_1=h\int_1^h \frac{s'(\rho)d\rho}{\rho(\rho^2-h^2)^{\frac{1}{2}}}. \tag{15.4}$$

Thus using Eq. (15.1)

$$\theta_1=2h\left[A\int_1^h \frac{d\rho}{\rho^2(\rho^2-h^2)^{\frac{1}{2}}} + B\int_1^h \frac{d\rho}{\rho[(\rho^2-h^2)(1-\rho^2)]^{\frac{1}{2}}}\right], \tag{15.5}$$

where by Eq. (14.5) $h=\sin\alpha_1$.

The first integral in Eq. (15.5) may be found in most standard tables and the second may be easily converted into a standard form, or, more directly, it may be evaluated by contour integration. Evaluating these

§§ Only rays that pass to right of axis will be considered in this discussion. For these θ and θ_1 are measured counterclockwise as one looks down on S. The direction is reversed for the rays passing to the left of the axis.

integrals Eq. (15.5) becomes simply

$$\theta_1=2A\left(\frac{\pi}{2}-\alpha_1\right)+B\pi=(A+B)\pi-2A\alpha_1. \quad (15.6)$$

Comparing this with Eq. (15.3), one sees that for the rays to be parallel in the cone K when developed one must have

$$\theta^*=(A+B)\pi \quad (15.7)$$

and

$$R=\sec\psi=2A. \quad (15.8)$$

From Fig. 7 the perpendicular distance between a ray and the reference ray is $R\sin\alpha$. The maximum width of the beam composed of rays passing on the same side of the axis is therefore R, the outer ray corresponding to $\alpha_1=\pi/2$.

Case I. $A+B=1$

For this special case, by Eq. (15.1), the surface S is given by the relationship

$$s(\rho)=(1-k)\rho+k\sin^{-1}\rho, \quad (15.9)$$

which defines the generalized Rinehart surfaces, see Table 1. Since

$$s'(\rho)=(1-k)+k(1-\rho^2)^{-\frac{1}{2}}, \quad (15.10)$$

one has that $s'(0)=1$ and $s'(1)=\infty$ (except for $k=0$). Thus, these surfaces all have a horizontal tangent at the axis, and, except for $k=0$ which corresponds to a plane surface, a vertical tangent at $\rho=1$.

For this one-parameter family of lenses, by Eq. (15.7), $\theta^*=\pi$ and the reference ray leaves diametrically opposite the source P. The rays passing to the left and right of the axis thus form a single beam. By Eq. (15.8) the cone K must have a slope angle given by

$$\psi=\sec^{-1}2(1-k). \quad (15.11)$$

Two important special cases already known are:

(1) RCA "R-2R" lens corresponding to $k=0$. By Eq. (15.11) the cone K is a 60° cone.

(2) Rinehart-Luneberg lens corresponding to $k=\frac{1}{2}$. For this choice of k, $\psi=\sec^{-1}1=0°$ and the cone K becomes a flat sheet.

Case II. $A+B>1$

The rays going to one side of the axis come out parallel to the reference ray, for which by Eq. (15.8),

$$\theta_1=\theta^*=(A+B)\pi>\pi. \quad (15.12)$$

The rays on the other side of the axis are parallel to a similar reference ray which is rotated through the same angle θ^* in the opposite direction. The angle between the reference rays and hence the angle between the beams is thus

$$2(A+B)\pi-2\pi=2(A+B-1)\pi. \quad (15.13)$$

Thus the beam is split, and one may obtain any desired angle between the beams. However, due to an unsymmetrical illumination of the aperture for each of the two beams, the secondary pattern for each beam will be distorted.

As an example, consider the following special case.

$$A=2,\ B=0.$$

By Eq. (15.7) each half-beam will be rotated 2π radians with respect to the source and hence will circle the axis and come out in the direction of the source. The two beams will thus form a single beam.

For this choice, by Eq. (15.1), S is a 60 degree cone, and by Eq. (15.8) the cone K is a cone with a slant angle of $\psi=\sec^{-1}4$. Thus the device consists of merely a cone on a cone. It may easily be verified that this lens may be obtained by halving the RCA "R-2R" lens by a plane through the axis and then joining up the edges of one of these halves. The paths of the geodesic rays as traced on the surfaces are clearly not affected by this modification and one may verify the above behavior of the rays.

16. GAUSSIAN CURVATURE OF THE OPTICAL SPACE

We shall generalize our treatment by letting the index of refraction be a function of s and θ. Thus the optical metric for our surface of revolution will be

$$d\zeta^2=n^2(s,\theta)[ds^2+\rho^2d\theta^2], \quad (16.1)$$

where as before, $d\zeta$ is the differential optical distance and s the arc length along the generating curve.

For a space in which the metric is given by

$$d\zeta^2=g_{11}(du^1)^2+g_{22}(du^2)^2, \quad (16.2)$$

the Gaussian curvature K is given by[11]

$$K=\frac{R_{1212}}{g}=-\frac{1}{2\sqrt{g}}\left[\frac{\partial}{\partial u^1}\left(\frac{1}{\sqrt{g}}\frac{\partial g_{22}}{\partial u^1}\right)+\frac{\partial}{\partial u^2}\left(\frac{1}{\sqrt{g}}\frac{\partial g_{11}}{\partial u^2}\right)\right], \quad (16.3)$$

where R_{1212} is a nonzero component of the Riemann tensor and

$$g=\begin{vmatrix} g_{11} & 0 \\ 0 & g_{22}\end{vmatrix}=g_{11}g_{22}. \quad (16.4)$$

If we let $u^1=s$ and $u^2=\theta$, and write

$$a=\sqrt{g_{11}},\quad b=\sqrt{g_{22}},\quad ab=\sqrt{g}, \quad (16.5)$$

Eq. (16.3) reduces at once to

$$K=-\frac{1}{ab}\left[\frac{\partial}{\partial s}\left(\frac{1}{a}\frac{\partial b}{\partial s}\right)+\frac{\partial}{\partial\theta}\left(\frac{1}{b}\frac{\partial a}{\partial\theta}\right)\right]. \quad (16.6)$$

[11] See I. S. Sokolnikoff, *Tensor Analysis* (John Wiley and Sons, Inc., New York, 1951), p. 169.

By a comparison of Eqs. (16.1) and (16.2)

$$a=n, \quad b=n\rho; \tag{16.7}$$

therefore,

$$K=\frac{-1}{n^2\rho}\left[\rho''+\frac{\partial}{\partial s}\left(\frac{\rho}{n}\frac{\partial n}{\partial s}\right)+\frac{1}{\rho}\frac{\partial}{\partial\theta}\left(\frac{1}{n}\frac{\partial n}{\partial\theta}\right)\right]$$

or

$$K=\frac{-1}{\rho n^2}\left[\rho''+\frac{\partial}{\partial s}\left(\rho\frac{\partial}{\partial s}\ln n\right)+\frac{1}{\rho}\frac{\partial^2}{\partial\theta^2}\ln n\right]. \tag{16.8}$$

In terms of the operator[11] [11]

$$\nabla^2=\frac{1}{\rho}\frac{\partial}{\partial s}\rho\frac{\partial}{\partial s}+\frac{1}{\rho^2}\frac{\partial^2}{\partial\theta^2}, \tag{16.9}$$

one may write that

$$K=-\frac{1}{n^2}\left(\frac{\rho''}{\rho}+\nabla^2\ln n\right). \tag{16.10}$$

For the special case of $n=1$ everywhere

$$K=-\rho''/\rho, \tag{16.11}$$

and for the special case of a flat guide $\rho=s=r$, and therefore

$$K=-n^{-2}\nabla^2\ln n. \tag{16.12}$$

If, as in the Rinehart-Luneberg lens, one has $s(\rho)$ given rather than $\rho(s)$, one replaces Eq. (16.11) by

$$K=-s''/(\rho s'^3). \tag{16.13}$$

Another modification of (16.11) is made by multiplying numerator and denominator by $2\rho'$; then

$$K=-2\rho'\rho''/2\rho\rho'=-d\rho'^2/d\rho^2=-(d/d\rho^2)(1/s'^2). \tag{16.14}$$

Applying Eq. (16.12) to the Luneberg lens for which $n(r)=(2-r^2)^{\frac{1}{2}}$, one finds that $K=4/(2-r^2)^3$. Thus $K=\frac{1}{2}$ at $r=0$ and $K=4$ at $r=1$. On the other hand, for the Rinehart analog of the Luneberg lens $s'=\frac{1}{2}[1+(1-\rho^2)^{\frac{1}{2}}]$, so that by Eq. (16.14) $K=4/[1+(1-\rho^2)^{\frac{1}{2}}]^3$. This agrees with the above K for the Luneberg lens if one applies the relation $1-r^2=(1-\rho^2)^{\frac{1}{2}}$ of Eq. (7.11).

17. TREATMENT OF GENERAL SURFACES BY MEANS OF DIFFERENTIAL INVARIANTS

One may express the relationship between equivalent guides by means of the differential invariants of the two-dimensional optical spaces associated with the guides.

If the optical metric defined on the mean surface is given by

$$d\zeta^2=g_{11}(du^1)^2+g_{22}(du^2)^2, \tag{17.1}$$

where u^1 and u^2 are orthogonal coordinates in the space, the differential invariants we wish to consider are

$$\Delta_1(f,h)=\frac{1}{g_{11}}\frac{\partial f}{\partial u^1}\frac{\partial h}{\partial u^1}+\frac{1}{g_{22}}\frac{\partial f}{\partial u^2}\frac{\partial h}{\partial u^2}, \tag{17.2}$$

$$\Delta_1 f=\frac{1}{g_{11}}\left(\frac{\partial f}{\partial u^1}\right)^2+\frac{1}{g_{22}}\left(\frac{\partial f}{\partial u^2}\right)^2=\Delta_1(f,f) \tag{17.3}$$

and for $g=g_{11}g_{22}$

$$\Delta_2 f=\frac{1}{\sqrt{g}}\left[\frac{\partial}{\partial u^1}\left(\frac{\sqrt{g}}{g_{11}}\frac{\partial f}{\partial u^1}\right)+\frac{\partial}{\partial u^2}\left(\frac{\sqrt{g}}{g_{22}}\frac{\partial f}{\partial u^2}\right)\right]. \tag{17.4}$$

When applied to a plane, $\Delta_1 f$ represents the square of the magnitude of the gradient of f, $\Delta_1(f,g)$ represents the dot product of the gradient of f and the gradient of h, and $\Delta_2 f$ the Laplacian of f.

Suppose one considers the equivalence between a curved guide with index of one, the optical metric for which in a geodesic polar coordinate system is[12]

$$d\zeta^2=ds^2+D^2(s,\phi)d\phi^2, \tag{17.5}$$

and a flat guide whose optical metric is

$$d\zeta'^2=n^2(r,\theta)[dr^2+r^2d\theta^2]. \tag{17.6}$$

Optical equivalence requires that the metrics apply to the same optical space and hence that the invariants of (17.2), (17.3), and (17.4) be equal for coordinates s, ϕ, and coordinates r, θ.

Letting $u^1=s$ and $u^2=\phi$, one has by Eq. (17.5) $g_{11}=1$ and $g_{22}=D^2$; therefore,

$$\Delta_1(f,h)=\frac{\partial f}{\partial s}\frac{\partial h}{\partial s}+\frac{1}{D^2}\frac{\partial f}{\partial\phi}\frac{\partial h}{\partial\phi}, \tag{17.7}$$

$$\Delta_1 f=\left(\frac{\partial f}{\partial s}\right)^2+\frac{1}{D^2}\left(\frac{\partial f}{\partial\phi}\right)^2, \tag{17.8}$$

and

$$\Delta_2 f=\frac{1}{D}\left[\frac{\partial}{\partial s}\left(D\frac{\partial f}{\partial s}\right)+\frac{\partial}{\partial\phi}\left(\frac{1}{D}\frac{\partial f}{\partial\phi}\right)\right]. \tag{17.9}$$

On the other hand, letting $u^1=r$ and $u^2=\theta$, one has, by Eq. (17.6), $g_{11}=n^2$ and $g_{22}=n^2r^2$; therefore

$$\Delta_1(f,h)=\frac{1}{n^2}\left[\frac{\partial f}{\partial r}\frac{\partial h}{\partial r}+\frac{1}{r^2}\frac{\partial f}{\partial\theta}\frac{\partial h}{\partial\theta}\right], \tag{17.10}$$

$$\Delta_1 f=\frac{1}{n^2}\left[\left(\frac{\partial f}{\partial r}\right)^2+\frac{1}{r^2}\left(\frac{\partial f}{\partial\theta}\right)^2\right], \tag{17.11}$$

and

$$\Delta_2 f=\frac{1}{n^2}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 f}{\partial\theta^2}\right]. \tag{17.12}$$

[11] [11] For a flat surface ($s=\rho$) this operator becomes the Laplacian.

[12] A geodesic polar coordinate system is defined in various texts on differential geometry. See, for example, D. J. Struik, *Lectures on Classical Differential Geometry* (Addison-Wesley Press, Cambridge, 1950), p. 137.

Case I

$n(r,\theta)$ is given to find $D^2(s,\phi)$.

(1) From Eqs. (17.11) and (17.8)

$$n^2\Delta_1 s=\left(\frac{\partial s}{\partial r}\right)^2+\frac{1}{r^2}\left(\frac{\partial s}{\partial \theta}\right)^2=n^2, \tag{17.13}$$

which one solves for $s=s(r,\theta)$.

(2) From Eqs. (17.10) and (17.7)

$$n^2\Delta_1(s,\phi)=\frac{\partial s}{\partial r}\frac{\partial \phi}{\partial r}+\frac{1}{r^2}\frac{\partial s}{\partial \theta}\frac{\partial \phi}{\partial \theta}=0, \tag{17.14}$$

which one solves for $\phi=\phi(r,\theta)$.

(3) From Eqs. (17.11) and (17.8)

$$\frac{1}{\Delta_1\phi}=D^2=\frac{n^2}{\left(\frac{\partial \phi}{\partial r}\right)^2+\frac{1}{r^2}\left(\frac{\partial \phi}{\partial \theta}\right)^2} \tag{17.15}$$

so that one obtains D^2 as a function of r and θ.

(4) One uses the solutions of Eqs. (17.13) and (17.14), i.e., $s=s(r,\theta)$ and $\phi=\phi(r,\theta)$, to obtain the functions

$$r=r(s,\phi), \quad \theta=\theta(s,\phi). \tag{17.16}$$

These are then substituted in Eq. (17.15) to obtain $D^2(s,\phi)$.

Case II

$D^2(s,\phi)$ is given, to find $n(r,\theta)$.

From Eqs. (17.9) and (17.12), letting $v=\ln r$,

$$\Delta_2 v=\frac{1}{D}\left[\frac{\partial}{\partial s}\left(D\frac{\partial v}{\partial s}\right)+\frac{\partial}{\partial \phi}\left(\frac{1}{D}\frac{\partial v}{\partial \phi}\right)\right]=0 \tag{17.17}$$

and

$$\Delta_2\theta=\frac{1}{D}\left[\frac{\partial}{\partial s}\left(D\frac{\partial \theta}{\partial s}\right)+\frac{\partial}{\partial \phi}\left(\frac{1}{D}\frac{\partial \theta}{\partial s}\right)\right]=0. \tag{17.18}$$

From Eqs. (17.7) and (17.10)

$$\Delta_1(v,\theta)=\frac{\partial v}{\partial s}\frac{\partial \theta}{\partial s}+\frac{1}{D^2}\frac{\partial v}{\partial \phi}\frac{\partial \theta}{\partial \phi}=0, \tag{17.19}$$

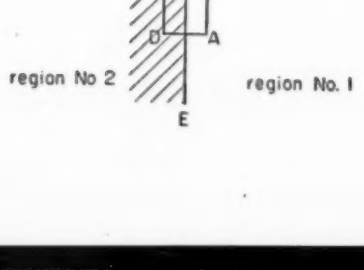


FIG. 8.

while from Eqs. (17.8) and (17.11)

$$\Delta_1 v=\left(\frac{\partial v}{\partial s}\right)^2+\frac{1}{D^2}\left(\frac{\partial v}{\partial \phi}\right)^2=\frac{1}{n^2r^2} \tag{17.20}$$

and

$$\Delta_1\theta=\left(\frac{\partial \theta}{\partial s}\right)^2+\frac{1}{D^2}\left(\frac{\partial \theta}{\partial \phi}\right)^2=\frac{1}{n^2r^2}. \tag{17.21}$$

Let λ be defined by the equation

$$\frac{\partial v}{\partial s}=\lambda\frac{1}{D}\frac{\partial \theta}{\partial \phi}; \tag{17.22}$$

then by Eq. (17.19)

$$\frac{1}{D}\frac{\partial v}{\partial \phi}=-\lambda\frac{\partial \theta}{\partial s}. \tag{17.23}$$

Since from Eqs. (17.20) and (17.21) $\Delta_1 v=\Delta_1\theta$, one sees that by Eqs. (17.22) and (17.23) $\lambda=1$.

One may, therefore, proceed as follows.

(1) One obtains a solution $f(s,\phi)$ of the differential equation

$$\frac{1}{D}\left[\frac{\partial}{\partial s}\left(D\frac{\partial f}{\partial s}\right)+\frac{\partial}{\partial \phi}\left(\frac{1}{D}\frac{\partial f}{\partial \phi}\right)\right]=0, \tag{17.24}$$

which by Eqs. (17.17) and (17.18) may be used as the expression either for $v=\ln r$ or for θ.

(2) One uses the equations, see Eqs. (17.22) and (17.23),

$$\frac{\partial v}{\partial s}=\frac{1}{D}\frac{\partial \theta}{\partial \phi}, \quad \frac{1}{D}\frac{\partial v}{\partial \phi}=-\frac{\partial \theta}{\partial s} \tag{17.25}$$

to obtain the other of the quantities v and θ.

(3) One obtains n^2 as a function of s and ϕ from either Eq. (17.20) or (17.21), i.e.,

$$n^2=\frac{1}{r^2\Delta_1\theta}=\frac{1}{r^2\Delta_1 v}. \tag{17.26}$$

(4) Since from the first two steps one has $v=v(s,\phi)$ and $\theta=\theta(s,\phi)$, one may obtain from them the inverse functions

$$s=s(v,\theta), \quad \phi=\phi(v,\theta). \tag{17.27}$$

From these we may express n^2 as a function of v and θ, or, since $v=\ln r$, as a function of r and θ. To make $n(r,\theta)$ physically meaningful we select the positive square root of n^2.

18. DISCONTINUOUS INDEX OF REFRACTION

If the index of refraction changes discontinuously at some boundary *EF*, see Fig. 8, one may form a small

rectangle $ABCD$ with AB in region 1, having an index n_1, and with CD in region 2, having an index n_2. Furthermore, at all regular points on the boundary the rectangle may be made small enough so that EF is approximately a straight line. One, therefore, may make AD and BC negligible in optical length compared to AB and CD. Thus two moving points at A and B will be separated at first by an optical distance n_1AB but on moving negligible distances AD and BC are separated by an optical distance of n_2CD that differs from n_1AB by a finite amount equal approximately to $AB(n_2-n_1)$.

One may easily convince oneself that there can be no such abrupt change in optical distances between moving points if the index is continuous. This means that all guides equivalent to a guide with a discontinuous index of refraction n must themselves have a discontinuous index. That is, one may not expect to find exact geodesic analogs of the usual optical system, since these employ dielectric lenses with boundaries across which the index of refraction changes abruptly.

BIBLIOGRAPHY

[1] H. B. DeVore and Harley Iams, RCA Rev. 9, 721–732 (1948).

[2] Ellis, Fine, and Reynolds, A Constant Thickness Achromatic Lens. Air Force Cambridge Research Laboratory Report (January, 1951), p. 14.

[3] Kunz, Brammer, and Johannesen, Final Report on Contract No. W28-099-ac-141, Experimental Phase. Final report on above contract between Case Institute of Technology and the Air Force Cambridge Research Laboratories (September 30, 1949).

[4] K. S. Kunz, Final Report on Contract No. AF(122)-92, Theoretical Phase. Final report on above contract between Case Institute of Technology and the Air Force Cambridge Research Laboratories (August 30, 1951).

[5] Harry Levy, Final Report on Study of Geodesics. Final report on Contract No. W-28-099-ac-154 between the University of Illinois and the Air Force Cambridge Research Laboratories (August 1, 1948).

[6] R. K. Luneberg, "Mathematical Theory of Optics" (mimeographed lecture notes), (Brown University, 1944), pp. 208–213.

[7] S. B. Myers, J. Appl. Phys. 18, 221–229 (1947).

[8] S. B. Myers, Parallel Plate Optics for Rapid Scanning. Radiation Laboratory Report 646, Massachusetts Institute of Technology (December 15, 1944).

[9] R. F. Rinehart, J. Appl. Phys. 19, 860–862 (1948).

[10] R. F. Rinehart and F. D. Parker, Parallel Plate Optics for Electrical Scanning. Final Report on Theoretical Phase of Contract W28-099-ac-141 between Case Institute of Technology and Air Force Cambridge Research Laboratory (June 30, 1948).

JOURNAL OF APPLIED PHYSICS VOLUME 25, NUMBER 5 MAY, 1954

Errors in Power Spectra Due to Finite Sample*

LEE M. SPETNER
Applied Physics Laboratory, The Johns Hopkins University, Silver Spring, Maryland

(Received September 11, 1953)

An estimate is made of the accuracy with which the "spectrum" of a finite sample of Gaussian noise approximates the spectrum of the original noise from which the sample was taken. The sample is considered to be spliced to form an endless loop, thereby making it a periodic function. The effect of this splice is also discussed.

INTRODUCTION

THE widespread use in statistical theory of the notion of power spectra of random processes stresses the importance of the experimental task of estimating these spectra. There is, however, a fundamental difficulty inherent in such an estimate, namely, that in order to measure accurately the power spectrum of a given random process it is necessary that there be available an infinite sample of this process with which to work. It is important that some estimate be made of the errors that arise because of the necessarily finite samples to which one is limited in practice. Tukey[1] has given, without proof, a clear-cut method by which one may estimate the accuracy of a spectrum derived from autocorrelation measurements. Also, Grenander and Rosenblatt[2] have stated four theorems, without proof, relevant to the estimate of the power spectrum of a discrete parameter random process.

We shall not enter here the argument as to which is better to measure, the autocorrelation function or the spectrum. But, rather, let us set for ourselves the task of estimating the error in the measured spectrum due only to the finiteness of the sample. Since in order to define a spectrum we require that the function exist for infinite time, we shall imagine that we splice the head of the sample to its tail, thereby forming an endless loop. This loop could then be played through a spectrum analyzer or a correlator as if it were an infinite sample. We shall then determine the accuracy with which the spectrum of the original noise can be estimated from that of this sample loop. We shall restrict ourselves to stationary Gaussian processes having spectra which are strongly continuous.†

It should be pointed out that the above splicing technique is very often applied in practice. The addi-

* This work was supported by the U. S. Navy Bureau of Ordnance under Contract NOrd 7386.

[1] J. W. Tukey, "Sampling Theory of Power Spectrum Estimates," Symposium on Application of Autocorrelation Analysis to Physical Problems (U. S. Office of Naval Research, June, 1949).

[2] U. Grenander and M. Rosenblatt, Proc. Natl. Acad. Sci. U. S. 38, 519 (1952).

† "Strongly continuous" here means that the magnitude of the spectrum does not change significantly in frequency intervals of the order of the reciprocal of the sample length.

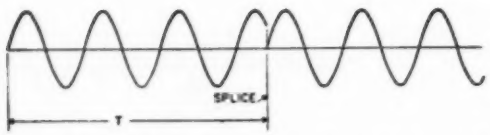


FIG. 1. Plot of $\sin_T 2\pi ft$. Note the discontinuity at the splice as the function repeats itself.

tional question then appears as to the effect of the discontinuity at the splice. It must be realized that the splice is an essential difficulty, and is not merely a peculiarity of an experimental procedure.

CASE OF SIMPLE SINUSOID

Before attacking the more difficult problem of random noise, let us examine the result of splicing a pure sine wave signal in the manner indicated above. We shall take, for a sine wave of frequency f, the portion contained in the time interval from 0 to T, and splice this into an endless loop. Figure 1 shows the resulting function, which is periodic with period T, and we shall call this function $\sin_T 2\pi ft$. Its Fourier series expansion is

$$\sin_T 2\pi ft = \tfrac{1}{2}a_0 + \sum_m [a_m \cos(2\pi mt/T) + b_m \sin(2\pi mt/T)] \quad (1)$$

where

$$a_m = 2T^{-1}\int_0^T \sin(2\pi ft)\cos(2\pi mt/T)dt$$

and

$$b_m = 2T^{-1}\int_0^T \sin(2\pi ft)\sin(2\pi mt/T)dt.$$

Integrating these we find

$$a_m = (1-\cos 2\pi fT)[(2\pi fT-2\pi m)^{-1} + (2\pi fT+2\pi m)^{-1}] \quad (2a)$$

$$b_m = \sin 2\pi fT[(2\pi fT-2\pi m)^{-1} - (2\pi fT+2\pi m)^{-1}]. \quad (2b)$$

Similarly for $\cos_T 2\pi ft$, we have

$$\cos_T 2\pi ft = \tfrac{1}{2}c_0 + \sum_m [c_m \cos(2\pi mt/T) + d_m \sin(2\pi mt/T)] \quad (3)$$

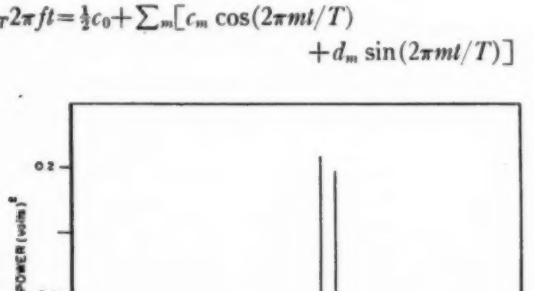


FIG. 2a. Power spectrum for a 4-cps sine wave chopped after 16 periods plus 180°, then spliced into an endless loop.

where

$$c_m = \sin 2\pi fT[(2\pi fT-2\pi m)^{-1} + (2\pi fT+2\pi m)^{-1}] \quad (4a)$$

$$d_m = -(1-\cos 2\pi fT)[(2\pi fT-2\pi m)^{-1} - (2\pi fT+2\pi m)^{-1}]. \quad (4b)$$

Notice that the spectrum that results is affected not only by the gross length of the sample but also to some degree by the phase at which the splice is made. In fact, as the sample length approaches an integral number of periods, i.e., as $T \to n/f$, we can see from

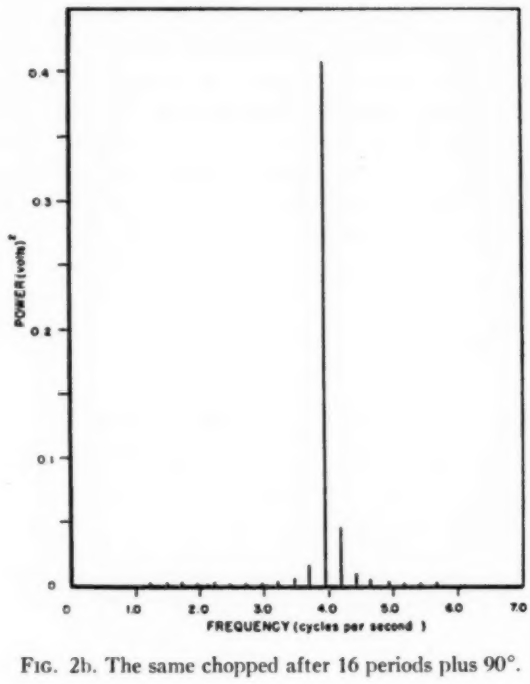


FIG. 2b. The same chopped after 16 periods plus 90°.

Eqs. (2) and (4) that

$$a_m \to 0 \text{ for all } m$$
$$b_m \to 0 \text{ for } m \neq n$$
$$b_n \to 1$$
$$c_m \to 0 \text{ for } m \neq n$$
$$c_n \to 1$$
$$d_m \to 0 \text{ for all } m.$$

Furthermore, one can see that as $T \to \infty$ for any phase at the splice, the frequency separation of the substantially nonvanishing lines will approach zero. Figure 2 shows a plot of $\frac{1}{2}(a_m^2+b_m^2)$ for a 4-cps sine wave, with two different gross lengths of sample, each with both a 90° splice and a 180° splice. From these one can see the tendency for the spectrum lines to bunch together at 4 cps as T increases.

SPECTRUM OF NOISE SAMPLE

Let us turn now to the problem of finding the spectrum of a random noise sample of length T. We shall

denote the noise by $I(t)$, choosing as our sample the portion of $I(t)$ contained between 0 and T. Splicing the ends to form an endless loop we produce a periodic function which we shall call $g(t)$, where

$$g(s\pm nT)=I(s), \quad 0\leq s<T, \quad n=0, 1, 2, \cdots. \quad (5)$$

Thus $g(t)$ is defined for all t $(-\infty<t<\infty)$.

Since $g(t)$ is periodic, its spectrum will contain discrete power impulses at integral multiples of $1/T$ on the frequency scale. The power spectrum of the original noise, $I(t)$, which we shall call $W(f)$ will be, in general, a continuous function of frequency. In order to facilitate the comparison of the measured spectrum with $W(f)$, let us define the function $w(f)$ to be T times the power of $g(t)$ at the frequency f, and if f is not a multiple of $(1/T)$ then we shall define $w(f)$ by means of the power impulse of $g(t)$ at the nearest integral multiple of $1/T$ that is less than f. (In practice, the measured spectrum will have some such smoothed out appearance as $w(f)$ because of the finite filter width in the spectrum analyzer.)

Writing $g(t)$ in its Fourier series form we have

$$g(t)=\tfrac{1}{2}A_0+\sum_m[A_m\cos(2\pi mt/T) + B_m\sin(2\pi mt/T)]. \quad (6)$$

Hence the power spectrum $w(f)$ can be written as

$$w(m/T)=\tfrac{1}{2}T(A_m^2+B_m^2) \quad (7)$$

where A_m and B_m are the Fourier coefficients of Eq. (6) In order to determine A_m and B_m we write $I(t)$ in its Fourier form after the fashion of Rice[3] (we may do this since we are assuming $I(t)$ to be Gaussian),

$$I(t)=\sum_k[\alpha_k\cos(2\pi f_kt)+\beta_k\sin(2\pi f_kt)], \quad (8)$$

where the α_k's and β_k's are independent Gaussian random variables with zero mean and variance of $W(f_k)\Delta f$, and $f_k=k\Delta f$. Equation (8) is to be considered in the limit as $\Delta f\to 0$.

When the sample of length T is chosen from $I(t)$ and spliced to form $g(t)$, each sinusoidal component of (8) is affected in the same manner as the sinusoids of Eqs. (1) and (3). Hence, using Eqs. (2), (4), and (8) and collecting terms we have for the Fourier coefficients of $g(t)$

$$A_m=\sum_k[\alpha_k\sin 2\pi f_kT+\beta_k(1-\cos 2\pi f_kT)] \times[(2\pi f_kT-2\pi m)^{-1}+(2\pi f_kT+2\pi m)^{-1}] \quad (9a)$$

$$B_m=\sum_k[-\alpha_k(1-\cos 2\pi f_kT)+\beta_k\sin 2\pi f_kT] \times[(2\pi f_kT-2\pi m)^{-1}-(2\pi f_kT+2\pi m)^{-1}]. \quad (9b)$$

From (9) we see that A_m and B_m are Gaussian random variables with zero mean. This is evident from the fact that α_k and β_k are Gaussian and A_m and B_m are each just a sum of many of these α's and β's. Furthermore, as shown in Appendix A, A_m and B_m are statistically independent. Consequently the probability

[3] S. O. Rice, Bell System Tech. J. **23**, 282 (1944).

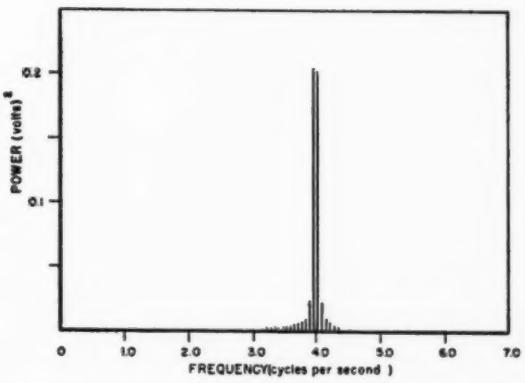


FIG. 2c. The same chopped after 64 periods plus 180°.

distribution of $w(m/T)$ can be calculated using Eq. (7). This calculation is performed in Appendix B and the result is

$$p(w)dw=\frac{dw}{T(\mathfrak{A}_m\mathfrak{B}_m)^{\frac{1}{2}}}\exp\left[-\frac{w}{2T}\left(\frac{1}{\mathfrak{A}_m}+\frac{1}{\mathfrak{B}_m}\right)\right] \times I_0\left[\frac{w}{2T}\left(\frac{1}{\mathfrak{A}_m}-\frac{1}{\mathfrak{B}_m}\right)\right] \quad (10)$$

where

$$w=w(m/T), \quad \mathfrak{A}_m=\langle A_m^2\rangle_{Av}, \quad \text{and} \quad \mathfrak{B}_m=\langle B_m^2\rangle_{Av}.$$

Next we must find expressions for the average values $\mathfrak{A}_m$ and $\mathfrak{B}_m$ in terms of $W(f)$. To accomplish this we square Eq. (9), average, and then pass to the limit as $\Delta f\approx df\to 0$. Using

$$\langle\alpha_k^2\rangle_{Av}=\langle\beta_k^2\rangle_{Av}=W(f_k)df$$

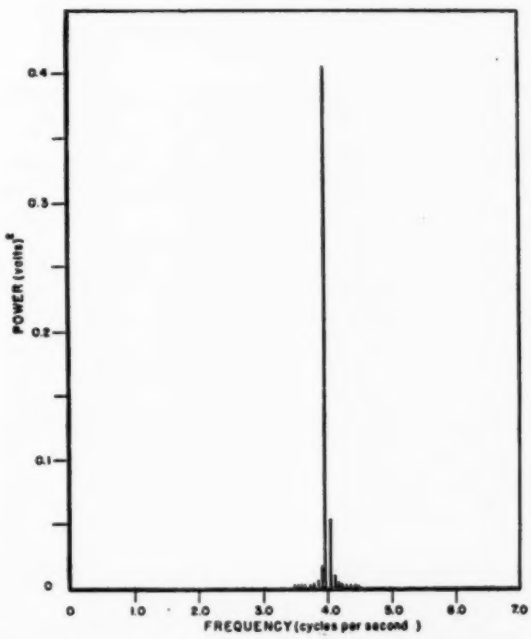


FIG. 2d. The same chopped after 64 periods plus 90°.

we obtain

$$\mathfrak{A}_m=\int_0^\infty W(f)\sin^2\pi fT[(\pi fT-\pi m)^{-1}+(\pi fT+\pi m)^{-1}]^2df \quad (11a)$$

$$\mathfrak{B}_m=\int_0^\infty W(f)\sin^2\pi fT[(\pi fT-\pi m)^{-1}-(\pi fT+\pi m)^{-1}]^2df. \quad (11b)$$

From Appendix C we find that these integrals can be reduced to

$$\mathfrak{A}_m=T^{-1}[W(m/T)\pm\delta_{ma}] \quad (12a)$$

$$\mathfrak{B}_m=T^{-1}[W(m/T)\pm\delta_{mb}] \quad (12b)$$

where δ_{ma} and δ_{mb} are both bounded by

$$|\delta_m|\leq 0.025Y_m+rX_m$$

where Y_m is the maximum value of $|W(f)-W(m/T)|$ considered as a function of f, X_m is the maximum value of $|W(f+m/T)-W(f-m/T)|$ considered as a function of f, and r is a number which depends not only upon m, but also on the frequency at which $W(f)$ becomes substantially flat. The value of r can be estimated from Fig. 3.

It is the magnitude of δ_m that represents the effect of the splice upon the spectrum measurement. The accuracy of the measurement of the ordinates of the spectrum will be poor at those frequencies which correspond to small values of m (i.e., when δ_m is large) simply because the average value of the measurement will have little relation to the ordinate of the true spectrum $W(m/T)$. For larger values of m, $\mathfrak{A}_m$ and $\mathfrak{B}_m$ will each more closely approach $T^{-1}W(m/T)$. When this occurs we replace $\mathfrak{A}_m$ and $\mathfrak{B}_m$ by this value in Eq. (10) and the probability distribution simplifies to

$$p(w)dw=e^{-w/W}dw/W \quad (13)$$

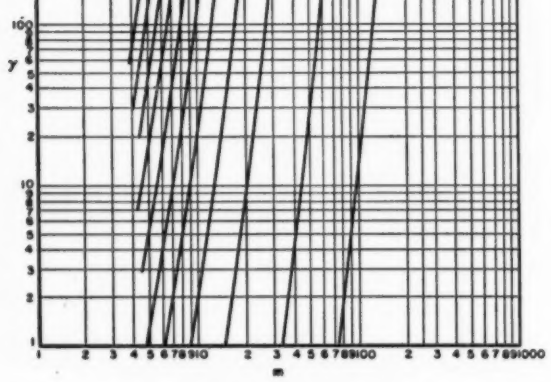


FIG. 3. r as a function of γ and m from Eq. (C.5). See text and Appendix C.

where we have written W for $W(m/T)$. Note that since $m=f_0T$, where f_0 is the frequency at which the ordinate is being considered, it can be large if the sample length T is large or if we only consider very high frequencies. In other words, the low frequency limit in the power spectrum measurement is determined by the sample length.

For example, let us say we are dealing with a spectrum that is substantially zero beyond 10 cps, and we have at our disposal a 5-minute sample from which we wish to determine the spectrum. Let us find the magnitude of δ at a frequency of 0.05 cps. We have the following:

$$T=300 \text{ sec}$$
$$\gamma=10/0.05=200$$
$$m=(0.05)(300)=15.$$

From Fig. 3 we find that $r\approx 0.035$. If we further say that the ordinate of the spectrum at 0.05 cps is of the same order of magnitude as the spectral maximum, we

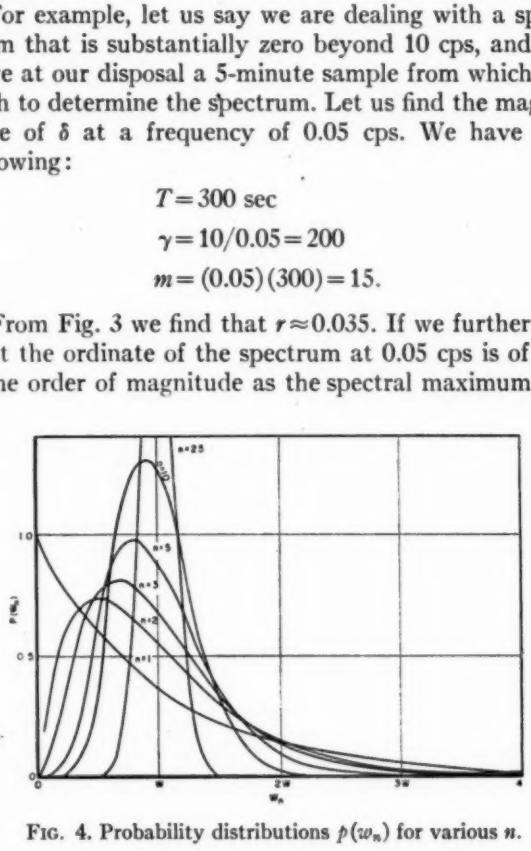


FIG. 4. Probability distributions $p(w_n)$ for various n.

can say that $\delta\approx 0.025W+0.035W=0.06W$, or

$$\mathfrak{A}_{15}\approx\mathfrak{B}_{15}\approx 0.00333W(0.05)$$

with about a six percent error. For frequencies lower than this the error will be greater, and for higher frequencies it will be smaller.

For small enough δ, Eq. (13) shows that the result of a measurement of $w(m/T)$ has a mean and standard deviation each equal to $W(m/T)$. Such a large standard deviation will not lead to reliable results, consequently the measuring procedure must be slightly altered. If one could obtain n independent samples of $I(t)$, each of length T, then for each value of m we could average all n measurements of $w(m/T)$, and we shall call this average w_n. It can be easily shown that this average has a probability distribution given by

$$p(w_n)dw_n=\frac{n^n}{(n-1)!}\left(\frac{w_n}{W}\right)^{n-1}e^{-nw_n/W}\frac{dw_n}{W}. \quad (14)$$

Plots of $p(w_n)$ for various n are shown in Fig. 4. The

mean of each of these distributions is W, and as n gets large one can see from Fig. 4 that the distribution narrows, and eventually approaches a Gaussian with standard deviation $Wn^{-\frac{1}{2}}$.

In actual practice, instead of obtaining n independent samples, one would use only one long sample (as long as is practical). He could then obtain the same results by averaging adjacent ordinates of the measured discrete spectrum, provided they are independent of each other. If the noise spectrum is substantially flat over frequency intervals of a few times the reciprocal of the sample length, this will be approximately correct. This averaging procedure is approximated in practice by the use of a wide filter in the spectrum analyzer.

Using Eq. (14) we can calculate probability bands about W in which w_n is likely to fall, and from these we can compute confidence regions about the measured value w_n in which W is likely to lie. For example, in Fig. 5, the region bounded by the two outer curves is the 90 percent confidence region, if we were to say that W is within this band we would be right 90 percent of the time. Similarly the two inner curves bound the 68.3 percent confidence region (this corresponds to plus and minus one standard deviation in the case of the Gaussian distribution).

CONCLUSION

We have thus indicated the accuracy and resolution one may obtain in estimating the power spectrum of noise using only a finite sample of it. The method discussed is that of the analog type as contrasted with Tukey's treatment[1] of the digital method. Our results are in agreement with his when the splice effect is negligible. Note that the number of his "degrees of freedom" are double our value of n for any particular point on the power spectrum curve since the sine and cosine parts of each frequency component are independent and form a degree of freedom each. Note further that Tukey's method gives twice as many degrees of freedom as we would estimate at each experimental point on the measured power spectrum, but this is quite understandable since in his case two adjacent points are not completely independent and in our case they are.‡ In the limiting case of flat noise the total number of degrees of freedom is approximately the same according to either method.

ACKNOWLEDGMENTS

The author wishes to thank Dr. C. I. Beard and Mr. I. Katz whose experimental work suggested the problem. He also wants to thank Mr. A. G. Carlton, Mr. J. H. Kuck, and Dr. A. J. Zmuda for many helpful discussions.

‡ His equivalent filter shapes are approximately triangular and they overlap quite a bit from point to point, whereas in our picture, the filter is assumed to be square in shape and nonoverlapping. (J. W. Tukey, private communication.)

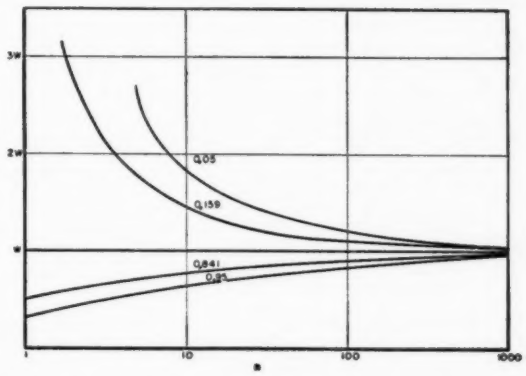


FIG. 5. Confidence bands of W for a measured value w. The index shown on the curves is the average percentage of time that W will be greater than the ordinate for any given n.

APPENDIX A. STATISTICAL INDEPENDENCE OF A_m AND B_m

The independence of A_m and B_m can easily be shown by writing down the joint probability of the k_{th} term in the summations of Eqs. (9a) and (9b). Using the fact that α_k and β_k are statistically independent and each Gaussian, one can show that the joint probability of

$$\mathfrak{a}_{mk}=[\alpha_k \sin 2\pi f_k T+\beta_k(1-\cos 2\pi f_k T)] \times[(2\pi f_k T-2\pi m)^{-1}+(2\pi f_k T+2\pi m)^{-1}]$$

and

$$\mathfrak{b}_{mk}=[-\alpha_k(1-\cos 2\pi f_k T)+\beta_k \sin 2\pi f_k T] \times[(2\pi f_k T-2\pi m)^{-1}-(2\pi f_k T+2\pi m)^{-1}]$$

is Gaussian with no cross term in the exponent, indicating that $\mathfrak{a}_{mk}$ and $\mathfrak{b}_{mk}$ are independent. This together with the fact that $\mathfrak{a}_{mk}$ and $\mathfrak{b}_{mj}$ are independent for $k\neq j$ (since α_k, β_k, α_j, and β_j are independent) it follows that A_m and B_m are also statistically independent.

APPENDIX B. THE PROBABILITY DISTRIBUTION $w(m/T)$

Using the result of Appendix A, we can compute the probability distribution of $w(m/T)$. Equation (7) states

$$w(m/T)=\tfrac{1}{2}T(A_m^2+B_m^2),$$

and using the fact that A_m and B_m are independent and each Gaussian, we can show easily that

$$p(w)dw=\frac{(\mathfrak{A}\mathfrak{B})^{-\frac{1}{2}}}{T}\exp\left[-\frac{w}{2T}\left(\frac{1}{\mathfrak{A}}+\frac{1}{\mathfrak{B}}\right)\right] \times I_0\left[\frac{w}{2T}\left(\frac{1}{\mathfrak{A}}-\frac{1}{\mathfrak{B}}\right)\right]dw,$$

where, for convenience we have omitted the subscripts m, and have let $w=w(m/T)$.

Note that if $\mathfrak{A}=\mathfrak{B}$, we have

$$p(w)dw=(\mathfrak{A}T)^{-1}\exp(-w/\mathfrak{A}T)dw.$$

APPENDIX C. PROOF OF EQS. (12a) AND (12b)

We shall evaluate here the integrals of Eqs. (11)

$$\mathfrak{A}_m=\int_0^\infty W(f)\sin^2\pi fT[(\pi fT-\pi m)^{-1}+(\pi fT+\pi m)^{-1}]^2df$$

$$\mathfrak{B}_m=\int_0^\infty W(f)\sin^2\pi fT[(\pi fT-\pi m)^{-1}-(\pi fT+\pi m)^{-1}]^2df.$$

Let us write for convenience

$$\mathfrak{A}_m=I_1+I_2+I_3$$

$$\mathfrak{B}_m=I_1-I_2+I_3$$

where

$$I_1=\int_0^\infty W(f)\frac{\sin^2\pi fT}{(\pi fT-\pi m)^2}df$$

$$I_2=2\int_0^\infty W(f)\frac{\sin^2\pi fT}{(\pi fT-\pi m)(\pi fT+\pi m)}df$$

$$I_3=\int_0^\infty W(f)\frac{\sin^2\pi fT}{(\pi fT+\pi m)^2}df.$$

If we assign $W(-f)=W(f)$, I_3 may be written as

$$I_3=\int_{-\infty}^0 W(f)\frac{\sin^2\pi fT}{(\pi fT-\pi m)^2}df.$$

Hence I_1 and I_3 may be combined to give

$$I_1+I_3=\int_{-\infty}^\infty W(f)\frac{\sin^2\pi fT}{(\pi fT-\pi m)^2}df.$$

Furthermore, the integrand in I_2 can be split using partial fractions to give

$$I_2=\frac{1}{\pi m}\int_0^\infty W(f)\frac{\sin^2\pi fT}{(\pi fT-\pi m)}df-\frac{1}{\pi m}\int_0^\infty W(f)\frac{\sin^2\pi fT}{(\pi fT+\pi m)}df.$$

As before, these two integrals may be combined to give

$$I_2=\frac{1}{\pi m}\int_{-\infty}^\infty W(f)\frac{\sin^2\pi fT}{(\pi fT-\pi m)}df.$$

Then, changing the variable of integration, we get

$$I_1+I_3=\int_{-\infty}^\infty W\left(\frac{m}{T}+f\right)\frac{\sin^2\pi fT}{(\pi fT)^2}df$$

and

$$I_2=\frac{1}{\pi m}\int_{-\infty}^\infty W\left(\frac{m}{T}+f\right)\frac{\sin^2\pi fT}{\pi fT}df. \qquad (C.1)$$

TABLE I. Values of $C_k=\pi^{-1}[\mathrm{Si}(2\pi k+2\pi)-\mathrm{Si}(2\pi k)]$. These are the coefficients of $W[(m+k)/T]$ in Eq. (C.2). One can obtain C_k for negative values of k from the relation $C_{-k}=C_{k-1}$.

k	C_k
0	0.4514
1	0.0236
2	0.0082
3	0.0042
4	0.0025
5	0.0017
6	0.0012
7	0.0009
8	0.0007
9	0.0006
10	0.0005
11	0.0004
12	0.0003
13	0.0003
14	0.0002
15	0.0002
16	0.0002
17	0.0002
18	0.0001
19	0.0001
20	0.0001

If we assume $W(f)$ to be substantially constant over integrals of length T^{-1} cps, we may approximate the integral I_1+I_3 by the series

$$I_1+I_3=\sum_{k=-\infty}^{\infty}W\left(\frac{m+k}{T}\right)\int_{k/T}^{(k+1)/T}\frac{\sin^2\pi fT}{(\pi fT)^2}df.$$

This integral can be immediately written as

$$\int_{k/T}^{(k+1)/T}\frac{\sin^2\pi fT}{(\pi fT)^2}df=\frac{1}{\pi T}[\mathrm{Si}(2\pi k+2\pi)-\mathrm{Si}(2\pi k)]$$

where

$$\mathrm{Si}(x)=\int_0^x t^{-1}\sin t\,dt.\S$$

If we let

$$C_k=\pi^{-1}[\mathrm{Si}(2\pi k+2\pi)-\mathrm{Si}(2\pi k)]$$

then we may write

$$I_1+I_3=\frac{1}{T}\sum_{k=-\infty}^{\infty}C_kW\left(\frac{m+k}{T}\right). \qquad (C.2)$$

Values of C_k are given in Table I. From this table we can see that if $W(f)$ is substantially constant in the neighborhood $(m-4)/T\leq f<(m+4)/T$ then

$$I_1+I_3=T^{-1}[0.975W(m/T)+e']$$

where

$$0<e'\leq 0.025[W(f)]_{\max},$$

§ For tables of this function see "Tables of Sine, Cosine, and Exponential Integrals," Vol. I and II; also "Tables of Sine and Cosine Integrals for Arguments from 10 to 100," National Bureau of Standards, Washington, D. C.

or better

$$I_1+I_3=T^{-1}[W(m/T)\pm 0.025Y_m]$$

where

$$Y_m \leq |W(f)-W(m/T)|_{\max}. \qquad (C.3)$$

Let us now estimate an upper limit for I_2. One must be careful in doing this since the integral

$$\int_0^\infty x^{-1}\sin^2 x dx$$

diverges. Rewriting Eq. (C.1), using the fact that

$$\frac{\sin^2\pi fT}{\pi fT}=-\frac{\sin^2\pi(-f)T}{\pi(-f)T}$$

we get

$$I_2=(\pi m)^{-1}\int_0^\infty \frac{\sin^2\pi fT}{\pi fT}\left[W\left(f+\frac{m}{T}\right)-W\left(f-\frac{m}{T}\right)\right]df.$$

Let us choose a γ large enough so that $W(f)$ is substantially constant for $f>\gamma m/T$. Then

$$I_2\approx(\pi m)^{-1}\int_0^{(1+\gamma)m/T}\frac{\sin^2\pi fT}{\pi fT}\left[W\left(f+\frac{m}{T}\right)-W\left(f-\frac{m}{T}\right)\right]df.$$

Let

$$X_m=\text{maximum of } |W(f+m/T)-W(f-m/T)|. \qquad (C.4)$$

Then

$$I_2\leq X_m(\pi m)^{-1}\int_0^{(1+\gamma)m/T}(\pi fT)^{-1}\sin^2\pi fT df.$$

But

$$\int_0^{(1+\gamma)m/T}(\pi fT)^{-1}\sin^2\pi fT df$$
$$=(2\pi T)^{-1}\{\ln[2\pi(1+\gamma)m]-\mathrm{Ci}[2\pi(1+\gamma)m]+0.577\}$$

where

$$\mathrm{Ci}(x)=\int_\infty^x t^{-1}\cos t dt.$$

Hence

$$I_2\leq X_m(2\pi^2 mT)^{-1}\{\ln[2\pi m(1+\gamma)]-\mathrm{Ci}[2\pi m(1+\gamma)]+0.577\}$$

or

$$I_2\leq rX_m/T$$

where

$$r=(2\pi^2 m)^{-1}\{\ln[2\pi m(1+\gamma)]+[2\pi m(1+\gamma)]^{-1}+0.577\} \qquad (C.5)$$

where we have used the fact that for large enough x

$$|\mathrm{Ci}(x)|<x^{-1}.$$

Figure 3 shows r as a function of m and γ.

Thus we may say that

$$\mathfrak{A}_m\approx T^{-1}[W(m/T)\pm\delta_m]$$
$$\mathfrak{B}_m\approx T^{-1}[W(m/T)\pm\delta_m]$$

where $|\delta_m|\leq 0.025Y_m+rX_m$, Y_m is given by Eq. (C.3), X_m is defined by (C.4) and r is to be obtained from Fig. 3.

We might point out here, that although γ is a rather vague quantity and hard to pin down precisely, this is not a serious disadvantage since a glance at Fig. 3 will show that r is a rather weak function of γ.

Approximate Emissivity Calculations for Polyatomic Molecules. I. CO_2*

S. S. PENNER
Guggenheim Jet Propulsion Center, California Institute of Technology, Pasadena, California
(Received March 9, 1953)†

Approximate emissivity calculations for CO_2 have been carried out, as a function of optical density, at 300 and at 600°K. The calculations involve the assumption that the rotational lines overlap extensively. This condition appears to be satisfied at total pressures above about 1 atmos. Comparison of the values calculated from spectroscopic data with the emissivities tabulated by Hottel and his collaborators shows satisfactory agreement. The analysis presented in this manuscript emphasizes the fact that it is possible to obtain reasonable estimates for the engineering emissivity without performing extensive analytical work, provided the physical principles are understood and the needed spectroscopic data are available.

I. INTRODUCTION

IN a series of recently published papers we have attempted to estimate the engineering emissivities of diatomic molecules by utilizing basic spectroscopic constants. Tractable methods of calculation have been developed for (a) complete overlapping between rotational lines[1] and for (b) separated rotational lines.[2] Although accurate theoretical calculations of gas emissivities from spectroscopic data for polyatomic molecules involve formidable computational difficulties, it is to be expected that approximate calculations can be carried out with relatively little effort by utilizing approximations similar to those introduced for the study of diatomic molecules.[1,2] For CO_2 it has been found that the total absorptivity at room temperature is substantially independent of pressure at pressures exceeding about 1 atmos,[3] thus indicating extensive overlapping between rotational lines. Accordingly, it is not unreasonable to base the present preliminary calculations on the assumption that extensive overlapping between rotational lines does occur. The results would be expected to apply, for example, at room temperature for total pressures in excess of 1 atmos and at 3000°K at pressures in excess of 11 atmos.[3] Since it appears quite likely that the actual range of validity of the results covers a larger range of total pressures, we shall not hesitate to show that calculated emissivities agree reasonably well with empirical data obtained[4] at 1 atmos at a temperature of 600°K as well as at 300°K. Perhaps the most important conclusion which can be derived from the present analysis is the statement that the analytical labor involved in making approximate emissivity calculations for a polyatomic molecule with overlapping rotational lines is trivial provided the important physical principles are understood and the needed spectroscopic data are available. It is only fair to note, however, that all of the needed integrated intensity measurements have not yet been performed for CO_2 in spite of the fact that several papers dealing with intensity measurements have been published recently.[5–7]

II. BASIC RELATIONS

Theoretical considerations of intensities for various transitions of the CO_2 molecule were summarized by Dennison a number of years ago.[8] We shall reproduce here the parts of the analysis which are useful for making approximate emissivity estimates on CO_2.

Frequencies (ν) and wave numbers (ω) corresponding to transitions between fixed energy levels are given by the Bohr frequency condition

$$\nu = c\omega = (\Delta W_V' + \Delta W_R')/h, \tag{1}$$

where $\Delta W_V'$ and $\Delta W_R'$ represent, respectively, the changes in vibrational and rotational energy corresponding to the frequency ν, h is Planck's constant, and c represents the velocity of light. According to the results of Dennison and Adel,[9,10] the rotational (W_R') and vibrational (W_V') energies for CO_2 are given, respectively, by the expressions

$$W_R' = hc(j^2+j-l^2)[0.3925-0.00058(n_1+\tfrac{1}{2}) +0.00045(n_2+1)-0.00307(n_3+\tfrac{1}{2})]+1.7hc(l^2-1) \tag{2}$$

and

$$W_V' = hc[1351.2(n_1+\tfrac{1}{2})+672.2(n_2+1)+2396.4(n_3+\tfrac{1}{2}) -0.3(n_1+\tfrac{1}{2})^2-1.3(n_2+1)^2-12.5(n_3+\tfrac{1}{2})^2 +5.7(n_1+\tfrac{1}{2})(n_2+1)-21.9(n_1+\tfrac{1}{2})(n_3+\tfrac{1}{2}) -11.0(n_2+1)(n_3+\tfrac{1}{2})]. \tag{3}$$

* Supported by the U. S. Office of Naval Research under Contract Nonr-220(03), NR 015 210.
† Failure to receive Referee's report and loss of the original manuscript caused unusual delay in publication time.
[1] (a) S. S. Penner, J. Appl. Phys. **21**, 685 (1950); (b) J. Appl. Mech. **18**, 53 (1951); (c) S. S. Penner and D. Weber, J. Appl. Phys. **22**, 1164 (1951).
[2] (a) Penner, Ostrander, and Tsien, J. Appl. Phys. **23**, 256 (1952); (b) S. S. Penner, J. Appl. Phys. **23**, 825 (1952).
[3] Holm, Weber, and Penner, J. Appl. Phys. **23**, 1283 (1952).
[4] W. H. McAdams, *Heat Transmission* (McGraw-Hill Book Company, Inc., New York, 1942), Chapter III by H. C. Hottel.
[5] A. M. Thorndike, J. Chem. Phys. **15**, 868 (1947).
[6] D. F. Eggers, Jr. and B. L. Crawford, Jr., J. Chem. Phys. **19**, 1554 (1951).
[7] Weber, Holm, and Penner, J. Chem. Phys. **20**, 1820 (1952).
[8] D. M. Dennison, Revs. Modern Phys. **3**, 280 (1931).
[9] D. M. Dennison, Revs. Modern Phys. **12**, 175 (1940).
[10] Slight corrections to the listed numerical values have been noted by W. S. Benedict and his collaborators. For the present purposes these corrections are negligibly small. Recent work on frequencies of lines belonging to CO_2 is described in the following papers: Benedict, Herman, and Silverman, J. Chem. Phys. **19**, 1325 (1951); Taylor, Benedict, and Strong, Progress Report on "Infrared Spectra of H_2O and CO_2 at 500°C," Contract Nonr 248-01, The Johns Hopkins University, March, 1952.

Here n_1, n_2, and n_3 are the vibrational quantum numbers associated with the fundamental vibration frequencies ν_1, ν_2, and ν_3, respectively. The quantum number l measures the angular rotation, in units of $h/2\pi$, which is associated with the degenerate (bending) ν_2 vibration. The rotational energy levels are identified by the rotational quantum number j.

The selection rules[9] for the transitions $n_1n_2{}^ln_3 \rightarrow n_1'n_2'^{l'}n_3'$ may be summarized as follows. For the perpendicular bands, Δn_2 odd, Δn_3 even, $\Delta l=\pm 1$. For the parallel bands, Δn_2 even, Δn_3 odd, $\Delta l=0$. The rotational selection rules are $\Delta j=\pm 1, 0$.

The amplitude factors[8] for given rotational transitions are[11]

$$\left(A_{j\ l}^{j\ l}\right)^2=\frac{l^2}{j(j+1)}, \tag{4}$$

$$\left(A_{j-1\ l}^{j\ l}\right)^2=\frac{j^2-l^2}{j(2j+1)}=\frac{2j-1}{2j+1}\left(A_{j\ l}^{j-1\ l}\right)^2, \tag{4a}$$

$$\left(A_{j\ l\mp 1}^{j\ l}\right)^2=\frac{(j\pm l)(j\mp l+1)}{4j(j+1)}, \quad l\neq 0, \tag{4b}$$

$$\left(A_{j\ 1}^{j\ 0}\right)^2=\tfrac{1}{2}, \tag{4c}$$

$$\left(A_{j-1\ l\mp 1}^{j\ l}\right)^2=\frac{(j\pm l)(j\pm l-1)}{4j(2j+1)}, \quad l\neq 0, \tag{4d}$$

$$\left(A_{j-1\ 1}^{j\ 0}\right)^2=\frac{j-1}{2(2j+1)}, \tag{4e}$$

$$\left(A_{j\ l\pm 1}^{j-1\ l}\right)^2=\frac{(j\pm l)(j\pm l+1)}{4j(2j-1)}, \quad l\neq 0, \tag{4f}$$

$$\left(A_{j\ 1}^{j-1\ 0}\right)^2=\frac{j+1}{2(2j-1)}. \tag{4g}$$

In Eqs. (4) to (4g) the convention has been adopted that the amplitude factor corresponding to the transition $j, l\rightarrow j', l'$ has been written as $\left(A_{j\ l}^{j'\ l'}\right)$.

The integrated absorption for a given transition from the lower-energy level $n_1n_2{}^ln_3$; j to the upper-energy level $n_1'n_2'^{l'}n_3'$; j' is designated as $S\equiv S(n_1n_2{}^ln_3;\ j\rightarrow n_1'n_2'^{l'}n_3';\ j')$ and is given[8] by the approximate relation

$$S=(8\pi^3\nu N_T/3hcQ_V'Q_R')\times\{\exp-[W_V'(n_1,n_2,n_3)+W_R'(j,l)]/kT\}\times g_{j'l'}\left(A_{j\ l}^{j'\ l'}\right)^2\beta^2[1-\exp(-h\nu/kT)]. \tag{5}$$

In Eq. (5), $\nu\equiv\nu(n_1n_2{}^ln_3;\ j\rightarrow n_1'n_2'^{l'}n_3';\ j')$ is the Bohr frequency corresponding to the indicated change in the quantum numbers;‡ N_T = total number of molecules per unit volume per unit pressure; $g_{j'l'}$ = statistical weight of the upper state with $g_{j'l'}=2j'+1$ for $l'=0$ and $g_{j'l'}=2(2j'+1)$ for $l'\neq 0$; $\beta\equiv\beta(n_1n_2{}^ln_3\rightarrow n_1'n_2'^{l'}n_3')$ is a factor which must be determined empirically and corresponds to the matrix element of the electric moment in the molecule associated with the indicated change in (vibrational) quantum numbers;‡ $Q_R'=\sum_j\sum_l g_{jl}\exp[-W_R'(j,l)/kT]$ = complete rotational partition function; $Q_V'=\sum_{n_1}\sum_{n_2}\sum_{n_3}\exp-[W_V'(n_1,n_2,n_3)/kT]$ = complete vibrational partition function.

The amplitude factors given in Eqs. (4) to (4g) are based on the assumption that vibration-rotation interactions have a negligibly small influence on amplitude factors. For this reason, the use of Eq. (5) involves somewhat cruder calculations than were employed in the calculations on diatomic molecules with nonoverlapping rotational lines.[2]

For the purposes of approximate radiant-heat transfer calculations, it is convenient to use the integrated absorption coefficient for a given band. Let

$$\alpha(n_1n_2{}^ln_3\rightarrow n_1'n_2'^{l'}n_3')=\sum_j\sum_{j'}S(n_1n_2{}^ln_3;\ j\rightarrow n_1'n_2'^{l'}n_3';\ j'). \tag{6}$$

For the parallel bands $\Delta l=0$ and

$$\sum_{j'}g_{j'l}\left(A_{j\ l}^{j'\ l}\right)^2=g_{jl}$$

according to the Burger and Dorgelo summation rules, which may be verified by use of Eqs. (4) to (4g) by utilizing the selection rules. Hence we obtain from Eqs. (5) and (6) the useful result

$$\alpha(n_1n_2{}^ln_3\rightarrow n_1'n_2'^{l'}n_3')=(8\pi^3\beta^2\nu'N_T/3hcQ_V)g_l\times\{\exp-[W_V(n_1,n_2,n_3,l)/kT]\}[1-\exp(-h\nu/kT)] \quad \text{with } g_l=1 \text{ for } l=0 \text{ and } g_l=2 \text{ for } l\neq 0, \tag{7}$$

where we have set

$$W_{V'}(n_1n_2n_3)+W_{R'}(j,l)=W_V(n_1,n_2,n_3,l)+W_R(j),$$

$$Q_V=\sum_{n_1,n_2,n_3,l}\exp(-W_V/kT), Q_V'Q_R'=Q_VQ_R \quad \text{with} \quad Q_R=\sum_j g_j\exp[-W(j)/kT],$$

and $g_{jl}=g_j\times g_l$. If the assumption is not made that $\nu\equiv\nu'\simeq\nu(n_1n_2{}^ln_3;\ 0\rightarrow n_1'n_2'^{l'}n_3';\ 0)$ a slight correction is obtained to Eq. (7).[12,13]

[11] The amplitude factors given in Eqs. (4) to (4g) are 4 times as large as those listed by Dennison (reference 8) because we are using a Fourier series in time of the form $\cos(2\pi\nu t)$ rather than $\exp(2\pi i\nu t)$.

‡ Although we shall not indicate explicitly the changes in the (vibrational) quantum numbers involved, both ν and β do, of course, vary from one vibration-rotation band to another.

[12] B. L. Crawford, Jr. and H. L. Dinsmore, J. Chem. Phys. **18**, 983, 1682 (1950).

[13] Eggers and Crawford (reference 6) have shown how to relate for CO_2 the matrix coefficients to the coefficients of Taylor series expansions of the potential energy and dipole moment about their respective equilibrium values.

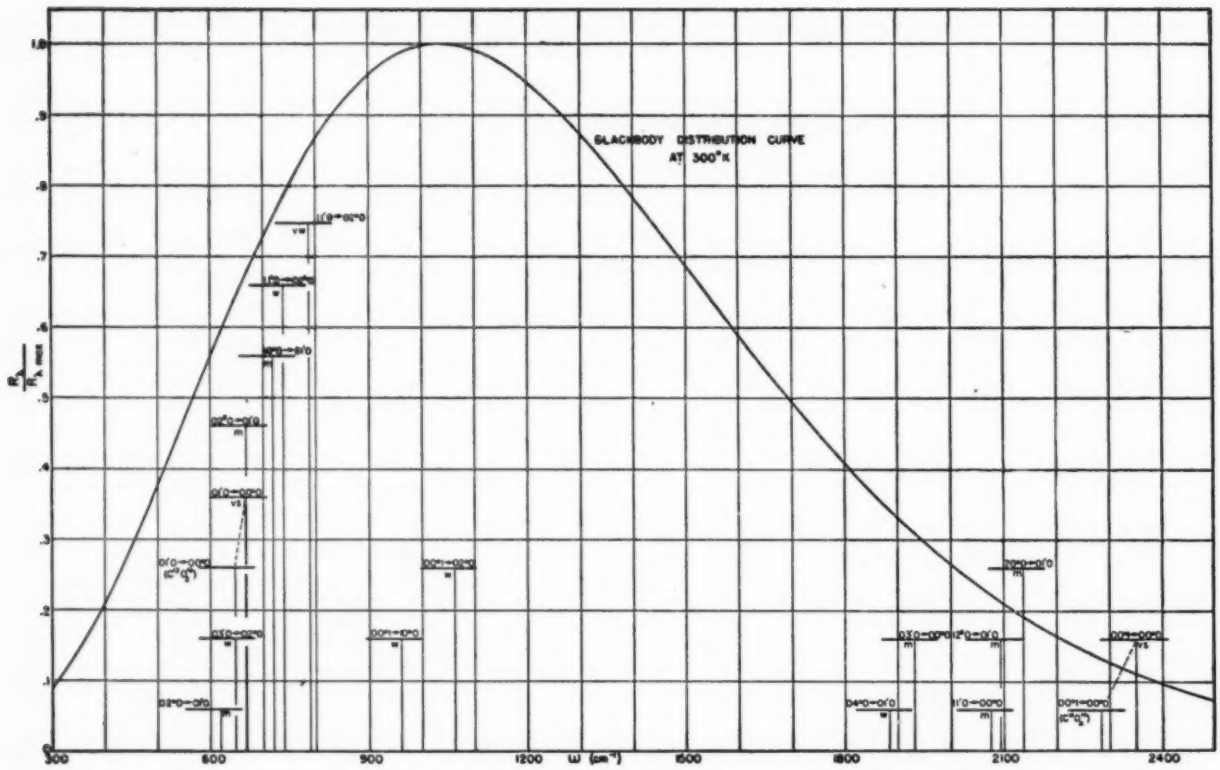


FIG. 1. Positions of centers of vibration-rotation bands, observed at room temperature, between 300 and 2400 cm^{-1}. The designations vs (very strong), s (strong), m (medium), w (weak), and vw (very weak) are those of Herzberg (reference 14). Also shown is the intensity ratio $R_\lambda/R_{\lambda\,\max}$ for a blackbody at 300°K.

For the perpendicular bands $\Delta l \neq 0$ and

$$\sum_{j',l'} g_{j'l'} \left(A_{j\ l}^{j'\ l'} \right)^2 = g_{jl},$$

$$\sum_{j'} g_{j'l'} \left(A_{j\ l}^{j'\ l\pm1} \right)^2 = \tfrac{1}{2} g_{jl}, \quad \sum_{j',l'} g_{j'l'} \left(A_{j\ 0}^{j'\ l'} \right)^2 = g_{jl}.$$

Hence

$$\alpha(n_1 n_2{}^l n_3 \rightarrow n_1' n_2{}'^{l+1} n_3') = (4\pi^3 \beta^2 \nu' N_T / 3hcQ_V) g_l \times \{\exp-[W_V(n_1,n_2,n_3,l)/kT]\}[1-\exp(-h\nu/kT)],$$

where $g_l = 1$ for $l = 0$ and $g_l = 2$ for $l \neq 0$, (8)

and

$$\alpha(n_1 n_2{}^l n_3 \rightarrow n_1' n_2{}'^{l-1} n_3') = (4\pi^3 \beta^2 \nu' N_T / 3hcQ_V) g_l \times \{\exp-[W_V(n_1,n_2,n_3,l)/kT]\}[1-\exp(-h\nu/kT)]$$

where $g_l = 1$ for $l = 0$ and $g_l = 2$ for $l \neq 0$. (9)

In using Eqs. (8) and (9) it should be noted particularly that the quantities β^2 are generally different for the transitions involving $l \rightarrow l+1$ and $l \rightarrow l-1$. The ratios of the β^2 can be calculated theoretically to the harmonic oscillator approximation for harmonic bands (see Appendix I for details).

In the same manner as for emissivity calculations on diatomic molecules with overlapping rotational lines,[1] we shall find Eq. (5) useful for the determination of "effective band widths" whereas Eqs. (7) to (9) lead directly to "average absorption coefficients" for vibration-rotation bands. For the calculation of S it may be convenient to combine Eqs. (4) to (4g), (5), and (7) to (9). For example, for the positive branch of the ν_2 fundamental of CO_2, we obtain

$$S \simeq [\omega(01^10; j \rightarrow 00^00; j-1)/\omega(01^10; 0 \rightarrow 00^00; 0)] \times (Q_R)^{-1} \times \alpha(01^10 \rightarrow 00^00) \times (j+1) \times [\exp-(W_R/kT)], \quad (5a)$$

where ω denotes a wave number.

III. APPROXIMATE EMISSIVITY CALCULATIONS AT 300°K

The positions and approximate intensities of the stronger vibration-rotation bands of CO_2 at 300°K are well known.[5–10,14] A first approximation to the effective band width is obtained by utilizing relations such as Eq. (5a) and defining the effective band width as the wave-number range for which S exceeds 10^{-3} of its maximum value.[1] Utilizing this definition of band width, which we shall refine presently, a summary of vibration-

[14] G. Herzberg, *Infrared and Raman Spectra of Polyatomic Molecules* (D. Van Nostrand Company, Inc., New York, 1951), Table 56 on p. 274. Herzberg's designations of band intensities as vs (very strong), s (strong), m (medium), w (weak), and vw (very weak) has been used in Figs. 1 to 3 to identify the approximate strengths of bands at room temperature.

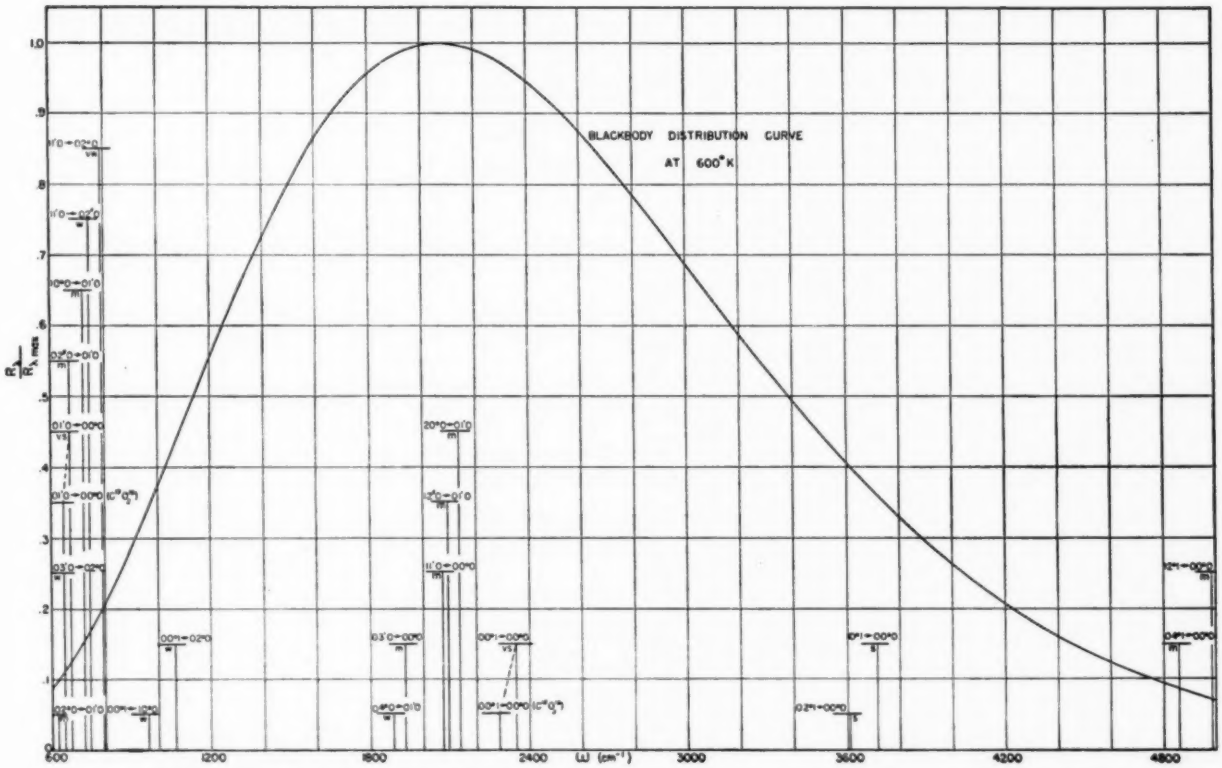


FIG. 2. Positions of centers of vibration-rotation bands, observed at room temperature, between 600 and 4800 cm^{-1}. Also shown is the intensity ratio $R_\lambda/R_{\lambda\ max}$ for a blackbody at 600°K.

rotation bands for CO_2 at 300°K has been drawn in Figs. 1 to 3. In these diagrams the abscissa is the wave number whereas the ordinate represents, in arbitrary units which vary from one band to another, the intensity. Integrated intensities at 300°K are given in Table I in so far as they are available.[7] For the strongest bands (ν_2 and ν_3 fundamentals) the contributions from the isotopic molecule $C^{13}O_2^{16}$ are also shown. This chemical species constitutes about 1.1 percent of all of the CO_2 present and makes contributions to the total emissivity which are not negligibly small for moderate and large optical densities of emitter. Also shown in Figs. 1 to 3 are the spectral intensities of radiation emitted by a blackbody at 300, 600, and 1500°K, again utilizing arbitrary scales along the ordinates, which are, however, consistent for any one of the given temperatures.

Reference to Fig. 1 shows that the principal contribution to the total radiant heat transfer at room temperature must occur from the spectral region around the ν_2 fundamental. Thus it is readily shown[1] that the "limiting emissivity" for CO_2 at 300°K is in the neighborhood of 0.4 with more than 90 percent of the total contributions made by the wave-number region extending from 551 to 830 cm^{-1}.

It is physically obvious that the engineering emissivity E at 300°K is determined by the $01^10 \rightarrow 00^00$ transition in the limit of zero optical density of the emitter. As was pointed out in an earlier publication,[1] the emissivity as a function of optical density (partial pressure p of emitter×optical pathlength L) for very small values of pL can be computed quite accurately either by utilizing a treatment applicable for nonoverlapping lines[2] or by using average absorption coefficients. The emissivity E_{667} of the band with center at 667 cm^{-1}, as calculated from Eq. (30) of reference 2a,

TABLE I. Integrated intensities for vibration-rotation bands of CO_2 (reference 7).

Band center (cm^{-1})	Transition	a(cm^{-2}-atmos^{-1})
5109	$00^00 \rightarrow 20^01$	0.426
4984	$00^00 \rightarrow 12^01$	1.01
4861	$00^00 \rightarrow 04^01$	0.272
3716	$00^00 \rightarrow 10^01$	42.3
3609	$00^00 \rightarrow 02^01$	28.5
2349	$00^00 \rightarrow 00^01$	2706
2137	$01^10 \rightarrow 20^00$	
2094	$01^10 \rightarrow 12^20$	0.147
2077	$00^00 \rightarrow 11^10$	
1933	$00^00 \rightarrow 03^10$	
1886	$01^10 \rightarrow 04^00$	0.083
721	$01^10 \rightarrow 10^00$	
668	$01^10 \rightarrow 02^20$	
667	$00^00 \rightarrow 01^10$	171.5
648	$02^00 \rightarrow 03^10$	
618	$01^10 \rightarrow 02^00$	
961	$10^00 \rightarrow 00^01$	0.0219[a]
1064	$02^00 \rightarrow 00^01$	0.0532[a]

[a] Unpublished data obtained by D. Weber.

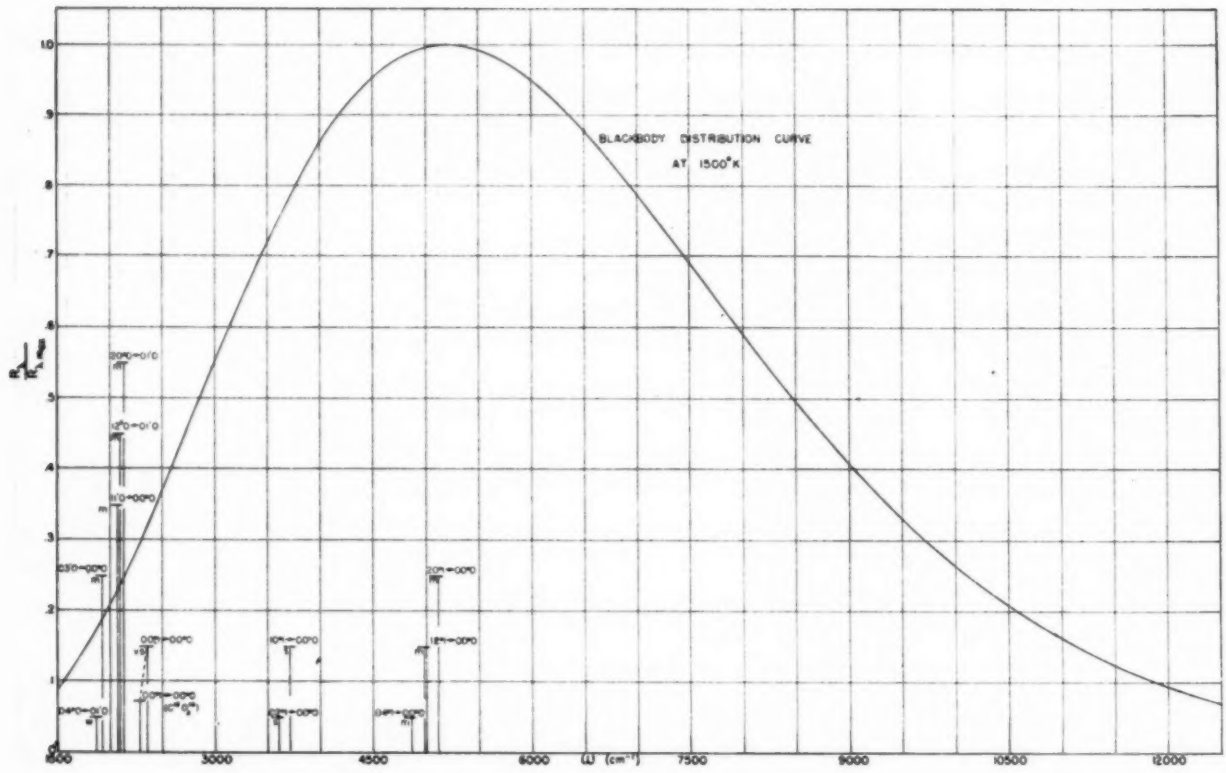


FIG. 3. Positions of centers of vibration-rotation bands observed at room temperature between 1500 and 12 000 cm^{-1}. Also shown is the intensity ratio $R_\lambda/R_{\lambda\ max}$ for a blackbody at 1500°K.

is plotted as a function of pL in Fig. 4 and labeled "approximation for nonoverlapping lines." The corresponding quantity obtained by using the average absorption coefficient[7] $\bar{P}X=(171\ cm^{-2}\ atmos^{-1})/(106\ cm^{-1})$ over the effective band width extending from 600 to 706 cm^{-1} gives results which are practically identical with those obtained for nonoverlapping rotational lines as pL is reduced below about 0.02 ft-atmos. As pL is increased Eq. (30) of reference (2a) rapidly fails to apply and yields excessively large values not only for E_{667} but also for the total engineering emissivity E. Needless to say, for $pL\leq 0.02$ ft-atmos, the numerical value of E_{667} is to be identified with the value of E.

As the optical density is increased to moderate values of pL, it becomes necessary first to include the $01^10\rightarrow 00^00$ transition for $C^{13}O_2^{16}$, which is assumed to represent 1.1 percent of the total CO_2, and for which the integrated intensity is roughly the same as for the $C^{12}O_2^{16}$ species. In treatments using average absorption coefficients it is, of course, necessary to add absorption coefficients rather than emissivities in regions in which partial overlapping between effective band widths occurs. Finally, the contributions made by the $10^00\rightarrow 01^10$ and $02^00\rightarrow 01^10$ bands must be included; both of these are designated by Herzberg[14] as being of "medium" intensity whereas the $03^10\rightarrow 02^00$ and $11^10\rightarrow 02^00$ bands are labeled "weak" and the $11^10\rightarrow 02^00$ band is said to be "very weak."

Strictly speaking, room-temperature emissivity calculations for CO_2 cannot be carried further because the integrated intensities for the weaker bands are not available. However, it is easily shown that results in fairly good agreement with Hottel's data[4] are obtained if reasonable estimates are made for the strongest of the bands which has not yet been included in determining E.

It is known from rough unpublished measurements[7] that the integrated intensity α_{721} for the $10^00\rightarrow 01^10$ band is of the order of 1 cm^{-2} $atmos^{-1}$. Emissivity calculations treating α_{721} as a variable parameter between 0 and 10 cm^{-2} $atmos^{-1}$ are shown in Fig. 4. The curves labeled $\alpha_{721}=0$, $\alpha_{721}=1$, and $\alpha_{721}=10$ with $\alpha_{667}=171$ correspond to the calculated values of E for the various assumed values of pL. In making these calculations the concept of the effective band width was refined in so far as the effective band width for the ν_2 fundamental was set equal to the wave-number range for which $SpL/0.7\geq 0.1$. This definition of band width for strong bands has the important advantage of making the band width a weak function of pL, in agreement with empirical measurements.[1b] The contributions to E made by the $00^01\rightarrow 10^00$ and $00^01\rightarrow 02^00$ bands are very small but have also been included. To these bands Weber assigns the following values at 300°K: $\alpha_{961}=0.0219\ cm^{-2}\ atmos^{-1}$, $\alpha_{1064}=0.0532\ cm^{-2}\ atmos^{-1}$.

In addition to the calculated emissivities, the results

of empirical measurements[4] are also shown in Fig. 4. For $pL \leq 0.06$ ft-atmos, the "empirical" data are really extrapolated results. Reference to Fig. 4 shows fair agreement between calculated and observed values of E for the reasonable values $\alpha_{667}=171$ and $\alpha_{721}=1$ cm^{-2} atmos^{-1}. At very small values of pL the calculated emissivities§ are larger than the extrapolated empirical data but must be considered to be more reliable since they depend only on the numerical value of α_{667}, which is known with fair accuracy.[5-7] Our ability to calculate E at very large values of pL is limited by the lack of adequate intensity data for the transitions $02^00 \rightarrow 01^10$, $11^10 \rightarrow 02^20$, and $11^10 \rightarrow 02^00$. However, it is clear that as pL is increased sufficiently above 3 ft-atmos, E must approach about 0.4 and will then increase only very slowly as exceedingly large values of pL are obtained.

The discussion of emissivity calculations on CO_2 at 300°K as a function of pL is mostly of academic interest. However, it is instructive in so far as it exemplifies the inherent simplicity of the analysis if adequate spectroscopic data are available.

IV. APPROXIMATE EMISSIVITY CALCULATIONS AT 600°K

It is evident from Fig. 2 that the major contributions to radiant heat transfer at 600°K are made by vibration-rotation bands in the spectral region between 890 and 2400 cm^{-1}. As pL goes to zero it is expected that the total emissivity E will become substantially equal to the emissivity of the intense ν_3 fundamental of $C^{12}O_2^{16}$ with appropriate corrections for contributions from the isotopic species $C^{13}O_2^{16}$. Although we are unable to complete the emissivity calculations at 600°K, because the needed spectroscopic data are not available, the engineering emissivity is estimated correctly for small values of pL. This result emphasizes the fact that relatively accurate emissivity calculations for polyatomic molecules can be performed without difficulty as soon as the necessary intensity measurements are available.

For the weaker bands, i.e., for all bands but the ν_3 fundamental, we shall define the effective band width as the wave-number region for which S exceeds about 10^{-3} of its maximum value. For the ν_3 fundamental we set the effective band width equal to the wave-number range for which $(SpL/0.7) \geq 0.1$.‖ The long wave-number limit of the ν_3 fundamental is set equal to 2410 cm^{-1}, i.e., 11 cm^{-1} beyond the band head. This choice allows for the many tails of rotational lines which "spill" across the band-head limit. We proceed by calculating separately the partial emissivities for selected wave-number regions.

A. Contributions of the $00^01 \rightarrow 10^00$ and $00^01 \rightarrow 02^00$ Bands

The criterion $(S/S_{\max}) \geq 10^{-3}$ for the weaker bands leads to band widths of the order of 140 cm^{-1} at 600°K. The integrated intensity for a given vibration-rotation band at the arbitrary temperature T, divided by the measured value at 300°K, is obtained by the use of Eqs. (7), (8), or (9). From these relations we find for the integrated intensity α of the band whose ground state is identified by the set of quantum numbers $n_1n_2^ln_3$ the result

$$\begin{aligned}\alpha(T°\mathrm{K})/\alpha(300°\mathrm{K}) &= (300/T)\\ &\times\{Q_V(300°\mathrm{K})\exp[W_V(00^00)/300k]\}\\ &\times\{Q_V(T°\mathrm{K})\times\exp[W_V(00^00)/kT]\}^{-1}\\ &\times\{\exp-[W_V(n_1n_2^ln_3)-W_V(00^00)]\\ &\times[(kT)^{-1}-(300k)^{-1}]\}[1-\exp(-h\nu/kT)]\\ &\times[1-\exp(-h\nu/300k)], \quad (10)\end{aligned}$$

where

$$\begin{aligned}Q_V(T)\exp[W_V(00^00)/kT] &\simeq [1-\exp(-1388hc/kT)]^{-1}\\ &\times[1-\exp(-667hc/kT)]^{-2}\\ &\times[1-\exp(-2349hc/kT)]^{-1}. \quad (11)\end{aligned}$$

The use of Eqs. (10) and (11) leads to the following results for the $00^01 \rightarrow 10^00$ and $00^01 \rightarrow 02^00$ bands, respectively: $\alpha_{961}(600°\mathrm{K}) \simeq 9.29\alpha_{961}(300°\mathrm{K}) = 0.204$ cm^{-2} atmos^{-1}; $\alpha_{1064}(600°\mathrm{K}) \simeq 7.22\alpha_{1064}(300°\mathrm{K}) = 0.385$ cm^{-2} atmos^{-1}. Hence the total contribution to the emissivity

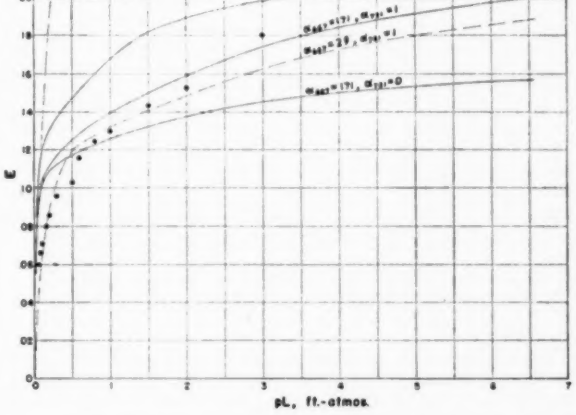


FIG. 4. Engineering emissivity E as a function of pL for CO_2 at 300°K. Integrated intensities α are expressed in cm^{-2} atmos^{-1}. The results based on empirical measurements (see reference 4) are shown as circles.

§ The dotted curve in Fig. 2 corresponds to $\alpha_{667}=29$ and $\alpha_{721}=1$ cm^{-2} atmos^{-1}. Although this curve represents a good fit of the empirical data, the result cannot be considered to be significant since α_{667} is known to be much larger than 29 cm^{-2} atmos^{-1}.

‖ Although the definitions of the effective band widths are somewhat arbitrary, it is easily shown that calculated values of E are quite insensitive to the chosen band widths. This result is caused by automatic choice of too small absorption coefficients for band widths which are too large, and conversely (compare reference 1a). Use of the quantity $SpL/0.7 \geq 0.1$ assures inclusion within the band width of the ν_3 fundamental of all of the rotational lines with average emissivities in excess of 0.1. Substantially the same numerical values for the effective band widths of the ν_3 fundamental are obtained if the lines with $SpL/0.7 \geq 10^{-1}n$ are included, where n is a number which does not differ from unity by more than a factor of three or four.

TABLE II. Absolute intensities at 300 and at 600°K for several vibration-rotation bands in the 5-micron region, based on intensity calculations using harmonic oscillator approximations (see Appendix I). The value 0.147 cm^{-2} $atmos^{-1}$ for the sum of the bands centered at 2137, 2094, and 2077 cm^{-1} served as standard to fix the absolute intensity scale.

Transition	α at 300°K (cm^{-2}-$atmos^{-1}$)	α at 600°K (cm^{-2}-$atmos^{-1}$)
$00^00 \rightarrow 11^10$	0.12	0.040
$01^10 \rightarrow 12^20$	0.020	0.032
$01^10 \rightarrow 20^00$	0.0045	0.0075
$01^10 \rightarrow 12^00$	0.00048	0.00079
$02^20 \rightarrow 13^30$	0.0012	0.0099
$03^30 \rightarrow 14^40$	0.000038	0.0020
$02^20 \rightarrow 21^10$	0.00019	0.0015
$10^00 \rightarrow 21^10$	0.00029	0.0026
$11^10 \rightarrow 30^00$	0.0000053	0.00026

obtained from the two bands under discussion is

$$E_{961}+E_{1064}=0.052[1-\exp(-1.40\times10^{-3}pL)]+0.017[1-\exp(-4.15\times10^{-3}pL)]+0.054[1-\exp(-2.75\times10^{-3}pL)], \quad (12)$$

where pL is expressed in cm-atmos.

B. Contributions Made by the Bands between 1800 and 2210 cm^{-1}

It is known from the recent work of Taylor, Benedict, and Strong[15] that an appreciable number of vibration-rotation bands, in addition to the bands shown in Fig. 1 in this region, are observable at 500°C.¶ Since all of these bands involve transitions to excited energy levels, their integrated intensities would be expected to increase as the temperature is raised. Rough absolute intensity estimates for several bands in the 5-micron region are given in Table II at 300 and 600°K. The relative intensity estimates were obtained by using the greatly oversimplified harmonic oscillator approximation described in Appendix I. Thus the contributions of both mechanical and of electrical anharmonicities to the intensity have been neglected.

As is evident from the experimental data shown in reference 15, a large number of vibration-rotation bands remain for which even rough intensity data are not available. For this reason no result can be given for the contributions made to the engineering emissivity by the vibration-rotation bands between 1800 and 2100 cm^{-1}.

TABLE III. Emissivities E_{2349} for $C^{12}O_2^{16}$ and $C^{13}O_2^{16}$ at 600°K as a function of optical density.*

pL, cm-atmos	pL, ft-atmos	$E_{2349}(C^{12}O_2^{16})$	$E_{2349}(C^{13}O_2^{16})$	E_{2349}(total)
0.1	0.0033	0.018	0.0005	0.019
0.5	0.0164	0.033	0.0016	0.035
1.0	0.033	0.037	0.0025	0.040
5.0	0.164	0.050	0.0070	0.057
15	0.492	0.051	0.011	0.062
50	1.64	0.052	0.011	0.063
100	3.28	0.053	0.012	0.065
200	6.56	0.055	0.012	0.067

* The tabulated values of $E_{2349}(C^{13}O_2^{16})$ correspond to the values which must be added to $E_{2349}(C^{12}O_2^{16})$ in order to obtain E_{2349}(total). They do not include emissivities in the (black) region of the ν_3 fundamental of $C^{13}O_2^{16}$ which is overlapped by the ν_3 fundamental of $C^{12}O_2^{16}$.

[15] Taylor, Benedict, and Strong, Progress Report on "Infrared Spectra of H_2O and CO_2 at 500°C," Contract Nonr 248-01, The Johns Hopkins University, Baltimore, March, 1952.

¶ Data of the type presented in reference 15 are valuable aids in performing emissivity calculations. For example, the spectroscopic emission records indicate, at least qualitatively, the vibration-rotation bands which must be included, at temperatures up to 773°K.

C. Contributions from the ν_3 Fundamental

From the measured integrated intensity[7] for the ν_3 fundamental at 300°K and by use of Eqs. (10) and (11) we find $\alpha_{2349}=1060$ cm^{-2} $atmos^{-1}$ at 600°K. Of the total integrated intensity per $(cm^{-1})\times(cm^{-1}$ $atmos^{-1}$ of CO_2) we associate 98.9 percent with $C^{12}O_2^{16}$ and 1.1 percent with $C^{13}O_2^{16}$. For the intense ν_3 fundamental the effective band width corresponds to the wavenumber range $\Delta\omega_{2349}$ for which $(SpL/0.7)\geq0.1$ with the band width of the isotopic species determined similarly. In general, the $C^{12}O_2^{16}$ band is black before the isotopic band makes appreciable contributions. The emissivities at 600°K, E_{2349}, calculated by the use of average absorption coefficients $\bar{P}=\alpha/\Delta\omega$, are summarized in Table III as a function of pL. Comparison of E_{2349}(total) with the empirical data[4] plotted in Fig. 5 shows that for $pL\leq0.033$ ft-atmos, E_{2349}(total) is practically identical with the engineering emissivity E. The result is in accord with expectations and shows

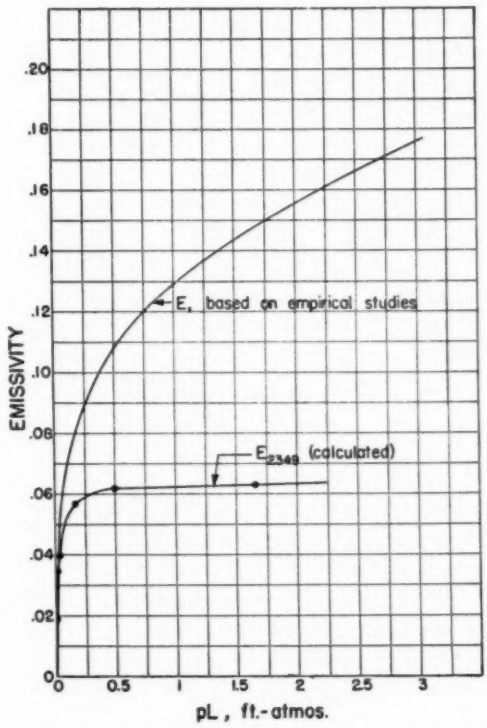


FIG. 5. Engineering emissivity E as a function of pL for CO_2 at 600°K, based on the empirical correlations of H. C. Hottel and R. B. Egbert. Also shown are the calculated contributions to the total emissivity made by the ν_3 fundamental.

clearly that useful emissivity data can be obtained with practically no analytical work once the physical principles are understood clearly.

D. Contribution of the $02^01\rightarrow00^00$ and $10^01\rightarrow00^00$ Bands

From the measured integrated intensities[7] we find $\alpha_{3609}=9.70$ cm^{-2} atmos^{-1} for the $02^01\rightarrow00^00$ band and $\alpha_{3716}=14.4$ cm^{-2} atmos^{-1} for the $10^01\rightarrow00^00$ band at 600°K. The effective band widths extend from 3539 cm^{-1} to 3679 cm^{-1} and from 3646 cm^{-1} to 3716 cm^{-1}, respectively. Using average absorption coefficients the following results are obtained:

$$E_{3609}+E_{3716}\simeq 0.0047[1-\exp(-6.92\times10^{-2}pL)]+0.0017[1-\exp(-1.72\times10^{-1}pL)]+0.0017[1-\exp(-1.03\times10^{-1}pL)]. \quad (13)$$

E. The Total (Engineering) Emissivity

The total (engineering) emissivity is obtained by adding the partial emissivities from selected spectral regions. Since estimates of partial emissivities were not included for all of the important vibration-rotation bands, no general expression for the engineering emissivity can be given although useful results have been obtained for small values of the optical density (compare Sec. IVC and Fig. 5).

APPENDIX I. RELATIVE INTENSITY CALCULATIONS FOR CO_2 USING HARMONIC OSCILLATOR APPROXIMATIONS**

It is well known that harmonic bands with nonzero matrix components are predicted even to the harmonic oscillator approximation. Calculations of this sort are useful for rough relative intensity estimates although mechanical as well as electrical anharmonicity corrections are neglected. Representative relative intensity estimates have been given, for example, by Benedict[15] and by Kaplan.[16] For the present applications it will be convenient to present a general equation for relative intensities of harmonic bands involving changes in the vibrational quantum number v_2.

The desired result is obtained readily by using the normalized wave functions for the isotropic plane oscillator and an integral involving associated Laguerre polynomials first obtained by Schrödinger.[17] In this manner it is easily shown that

$$[\beta(n_1n_2{}^ln_3\rightarrow n_1n_2{}'^{l'}n_3)]^2/[\beta(n_10^0n_3\rightarrow n_11^1n_3)]^2 = R(n_2{}^l\rightarrow n_2{}'^{l'})=[k!k'!/(k+l)!]\times(p!)^2\left\{\sum_{\tau=0}^{\leq k,k'}\binom{p-l}{k-\tau}\binom{p-l'}{k'-\tau}\binom{-p-1}{\tau}\right\}^2, \quad (A1)$$

where $k=(1/2)(n_2-l)$, $k'=(1/2)(n_2'-l')$, $p=(1/2)(l+l'+1)$, and $\binom{n}{k}$ denotes a binomial coefficient.

By combining Eqs. (7) to (9) with Eq. (A1) the following useful result is obtained:

$$\alpha(n_1n_2{}^ln_3\rightarrow n_1n_2{}'^{l'}n_3)/\alpha(n_10^0n_3\rightarrow n_11^1n_3)=\delta(n_2{}^l\rightarrow n_2{}'^{l'}) = g_lR(n_2{}^l\rightarrow n_2{}'^{l'})[\nu(n_1n_2{}^ln_3\rightarrow n_1n_2{}'^{l'}n_3)/\nu(n_10^0n_3\rightarrow n_11^1n_3)]\{\exp-[W_V(n_1,n_2,n_3,l)-W_V(n_1,0,n_3,0)]/kT\}\{1-\exp[-h\nu(n_2{}^l\rightarrow n_2{}'^{l'})/kT]\}\times\{1-\exp[-h\nu(0^0\rightarrow1^1)/kT]\}^{-1}. \quad (A2)$$

Numerical values of $R(n_2{}^l\rightarrow n_2{}'^{l'})$ calculated from Eq. (A1) are identical with Benedict's estimates,[15] which utilized data given by Shaffer.[18]

Following Benedict we shall assume that resonance between $2v_2$ and v_1 is exact, an approximation which will not change the order of magnitude of the calculated results although it is not in accord with observed infrared and Raman intensities. By utilizing Eqs. (A1) and (A2) relative intensities have been computed for a number of bands in the 5μ region at 300 and at 600°K. The results are summarized in Table (A-I).

TABLE (A-I). Relative intensities for harmonic bands in the 5μ region based on harmonic oscillator wave functions.

Transition	$g_lR(n_1n_2{}^l0\rightarrow n_1'n_2'{}^{l'}0)$	$\delta(n_1n_2{}^l0\rightarrow n_1'n_2'{}^{l'}0)$ at 300°K	$\delta(n_1n_2{}^l0\rightarrow n_1'n_2'{}^{l'}0)$ at 600°K
$00^00\rightarrow11^10$	1	1	1
$01^10\rightarrow12^20$	4	0.164	0.810
$01^10\rightarrow20^00$	0.9[a]	0.0376	0.187
$01^10\rightarrow12^00$	0.1[a]	0.00395	0.0196
$02^20\rightarrow13^30$	6	0.0101	0.248
$03^30\rightarrow14^40$	8	0.000315	0.0507
$02^20\rightarrow21^10$	0.9[a]	0.00155	0.0382
$10^00\rightarrow21^10$	1.8[a]	0.00238	0.0656
$11^10\rightarrow30^00$	0.9[a]	0.0000438	0.00638

[a] Adjusted according to Benedict to concentrate most of the intensity of resonating bands in the highest frequency member (reference 15).

** The author is indebted to Dr. W. S. Benedict and to Dr. L. D. Kaplan for helpful correspondence concerning intensity estimates for harmonic bands of CO_2.

[16] L. D. Kaplan, J. Chem. Phys. **18**, 186 (1950).

[17] E. Schrödinger, Ann. Physik **80**, 483 (1926).

[18] W. H. Shaffer, Revs. Modern Phys. **16**, 245 (1944).

Cross Section of Pulled Crystals

ROBERT G. POHL
Research Department, The Rauland Corporation, Chicago 41, Illinois
(Received August 21, 1953)

An equation for the radius of a crystal grown by the Czochralski technique is derived by considering the mechanical and thermal equilibrium at the liquid-solid interface. The results agree qualitatively with experimentally known functional dependencies.

IN growing a crystal by the Czochralski technique,[1,2] which consists of slowly withdrawing a seed crystal from a melt maintained near the melting point of the material, the size of the crystal is found to depend on the melt temperature and the velocity of crystal growth. In this note an equation is derived relating the crystal dimensions to the melt temperature, the growth velocity, the surface tension, and other physical constants. Then, from measurements on single crystals of germanium grown with varying temperatures and growth rates, values of the surface tension of germanium are calculated using the derived relation, and compared with directly measured values.

First, there is a relation between the true crystal growth velocity V, and the velocity of seed crystal withdrawal V', that must be demonstrated. The surface of the melt will be lowered as the crystal is grown; this velocity of lowering v, adds to the velocity with which the seed crystal is withdrawn in determining the crystal size. If x represents distance measured vertically along the axis of the crucible, this statement is represented by the equation

$$V(x)=v(x)+V'. \tag{1}$$

The relative areas of the crucible and of the growing crystal determine v, and if the crucible is of nonuniform cross section, v will be a function of the level height in the crucible. If $Y(x)$ is the radius of the crucible at a vertical height x measured from the bottom of the crucible, and $A(x)$ is the area of the liquid-solid interface, by applying conservation of mass to the system the following relation is easily obtained:

$$v(x)=\frac{V'}{\dfrac{\rho_L\pi Y(x)^2}{\rho_S A(x)}-1}, \tag{2}$$

where ρ_L and ρ_S are the densities in the liquid and solid phases, respectively. This correction term can be significant in practice because the crystal size may be comparable to that of the crucible it is grown from.

Next, consider the dynamical relations existing at the crystal-melt interface. Assuming the interface is planar, the seed, acting through the mechanisms of surface tension and liquid-solid cohesive forces, will lift the melt a small distance "a" above the liquid level. The calculation of these forces requires some knowledge of the shape of this small element of liquid which can be obtained by visual inspection of the region during the actual growth of the crystal. For the particular case of crystals grown from melts of germanium, an approximate shape of the system is represented in Fig. 1, wherein typical values for the dimensions are $r=1$ cm, $a\sim0.2$ cm, $b\sim0.2$ cm. The forces acting upward on this element of liquid, which to first order can be considered as a right circular cylinder, are the surface tension at the periphery and the liquid-solid cohesive force. The tangential shape of the surface of the element, where it meets the melt in the crucible, causes all the surface tension forces at the lower surface to be horizontal so that the only forces acting downward on the element are its weight and the liquid-liquid cohesive forces. The discussion below will proceed with the following simplifying assumptions: 1. the liquid-liquid cohesive force is negligible; 2. the solid-liquid cohesive force is small; 3. the dimension b is small compared to r so that the weight of the liquid lifted above the level of the melt can be approximated by $\rho_L g\pi r^2 a$. Then, by equating the forces upward to those acting downward, the distance "a" is found to be for equilibrium

$$a=\frac{2\gamma}{\rho_L r g}, \tag{3}$$

in which γ is the surface tension, and g is the acceleration of gravity.

Now the relation between V, r, and melt temperature can be found by considering the heat transfer processes occurring at the interface. Let W=heat conducted away from the interface by the crystal per unit time,

$$W=K_S\pi r^2\frac{\Delta T}{\Delta x}, \tag{4}$$

where K_S is the thermal conductivity of the solid, and $\Delta T/\Delta x$ is the temperature gradient at the interface. The heat arriving at the interface is that liberated by crystallization and that conducted from the melt. The heat loss by radiation from this small region of thickness a will be neglected so that the total heat brought to

[1] J. Czochralski, Z. physik. Chem. 92, 219 (1918).
[2] H. E. Buckley, *Crystal Growth* (John Wiley and Sons, Inc., New York, 1951), p. 83.

the interface per second is

$$K_L\pi r^2\frac{T-T_0}{a}+L\rho_L\pi r^2V, \tag{5}$$

wherein K_L is the thermal conductivity of the liquid phase, T is the melt temperature, T_0 is the melting point of the material, and L is the heat of fusion. By equating (4) and (5), substituting in the value of a from (3), and dividing through by πr^2, the desired relation is obtained,

$$\frac{K_L\rho_L g}{2\gamma}(T-T_0)r+L\rho_L V=K_S\frac{\Delta T}{\Delta x}\equiv W'. \tag{6}$$

This relation shows clearly the primary dependence of the crystal radius, and hence its cross-sectional area, on the various variables. For example, with V and W' constant, and neglecting supercooling,

$$r\sim\frac{1}{T-T_0}$$

or, with $(T-T_0)$ and W' constant,

$$r\sim W'-L\rho_L V,$$

so that as V increases, either by increasing the seed withdrawal rate or decreasing the crucible radius, r decreases. Or, with $(T-T_0)$ and r constant the growth velocity

$$V\sim K_S.$$

All of the above conclusions qualitatively agree with known experimental dependencies.[3] The above derivation assumes that V is less than the pertinent maximum Kristallisationgeschwindigkeit,[3] otherwise a true crystal will not be grown. The effects of supercooling, constitutional supercooling, and the incubation distance for cellular substructure[4] and of preferential crystal growth velocities have been neglected completely in the preceding derivation.

An explicit relation for r as a function of crucible shape can be obtained by substituting Eq. (2) in (6). Assuming $\rho_L=\rho_S$ and defining $C_0=K_L\rho_L g/2\gamma$, then,

$$r^3\left[\frac{C_0(T-T_0)}{V^2}\right]-r^2\left[\frac{W'}{V^2}\right]-r[C_0(T-T_0)]+[W'-L\rho_L V']=0,$$

the solution of which can be obtained from the cubic formula.

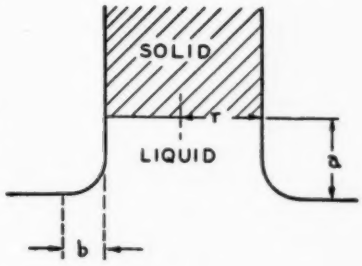


FIG. 1. Schematic representation of the element of liquid lifted by the seed crystal.

An attempt was made to verify (6) by varying the pulling rate from 6 to 2 inches per hour in five zones, and the melt temperature over a 5°C interval of a 100 gram single crystal of n-type germanium (having a resistivity of 3.5±0.5 ohm cm over the entire length). By measuring the radius of the crystal in each of the zones and solving the set of five simultaneous equations thus obtained for the unknown coefficients in (6), and assuming that $K_L\sim 0.1$ cal/sec cm°C and $g=980$ cm²/sec, the value obtained for the surface tension at the melting point is $\gamma\sim 340$ dyne/cm. This is to be compared with the value $\gamma=600$ dyne/cm recently obtained[5] by a more direct method. There are three uncertainties involved in the above calculation of γ from (6), the effect of crucible shape was neglected, the value for K_L was obtained by a perhaps unwarranted extrapolation of the observed[6] temperature variation of K_S, and in the solution of the set of simultaneous equations, the differences in crystal radii in the various zones rather than the radii themselves enter which reduces the accuracy of the results. Possibly, the value of K_L may be higher than that assumed, which would tend to reduce the discrepancy. Further experimental work would be highly desirable to provide additional quantitative tests of Eq. (6).

[3] G. Tammann, *Kristallisieren und Schmelzen* (Verlag J. A. Barth, Leipzig, 1903).
[4] Tiller, Jackson, Rutter, and Chalmers, Acta Metallurgica **1**, 428 (1953).
[5] P. H. Keck and W. Van Horn, Phys. Rev. **91**, 512 (1953).
[6] A. Grieco and H. C. Montgomery, Phys. Rev. **86**, 570 (1952).

Some New Magneto-Mechanical Torsion Experiments

A. W. COCHARDT
Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania
(Received September 3, 1953)

The internal damping of ferromagnetic wires was measured as a function of either a superimposed static stress or an alternating magnetic field. Similar to a static magnetic field, a static stress affects the damping considerably. For example, the logarithmic decrement of a Fe-40 percent Co wire in torsional vibration is reduced by a factor of 33 when a heavy weight is suspended on the wire. The effect of an alternating magnetic field on the damping was studied by varying the frequency of the field. It is found that the damping reaches a strong minimum at a field frequency of the order of magnitude of 20 cycles/sec. Magnetic hysteresis loss measurements were made which explain this phenomenon in terms of magneto-mechanical hysteresis.

INTRODUCTION

IT is well known that ferromagnetic materials exhibit relatively large internal damping or internal friction. This is mainly due to an effect, called magneto-mechanical hysteresis, which was discovered by Becker and Kornetzki 20 years ago.[1] According to this effect, ferromagnetic domains, when under the influence of an external mechanical stress, align themselves in a similar fashion as they do in a magnetic field. Since the alignment of the domains is partly an irreversible process, energy is dissipated in realigning domains during a mechanical stress-strain cycle. The area of the magnetic hysteresis loop is thereby closely related to the area of the mechanical stress-strain hysteresis loop. From these considerations it is apparent that the damping of mechanical vibrations in ferromagnetic materials depends, to a large extent, on the magnetic hysteresis loop.

Most of the previous work on magneto-mechanical hysteresis has dealt with the effect of a static magnetic field on the damping.[1–3] Evidently, the damping decreases when the material is subjected to a static magnetic field. The domains are then fixed in the direction of the field and are not able to move under the influence of the vibrational stress. Thus, virtually no energy is dissipated, and the damping must be small. In the present paper, the effects of a static stress and an alternating magnetic field are investigated. The effect of a superimposed static stress is of practical importance since machine members that exhibit magneto-mechanical damping[4] are often subjected to large centrifugal forces. The effect of an alternating magnetic field, on the other hand, is more of theoretical interest. While it is known that a 60-cycle field reduces the damping more than a static magnetic field,[5] a systematic study of this phenomenon has not yet been made. Nor to the author's knowledge has this effect been explained.

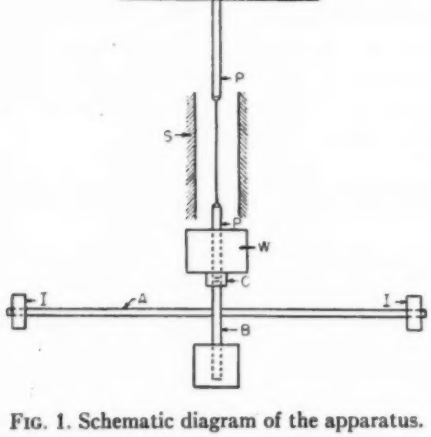


FIG. 1. Schematic diagram of the apparatus.

TABLE I. Composition and dimension of the test wires.

No.	%Fe	%Cr	%Co	Length (in)	Diameter (mil)
1	87	13	···	12.0	52.5
2	60	···	40	11.5	59.5
3	99.98	···	···	12.0	51.0
4	65	···	35	12.0	59.0

EFFECT OF STATIC STRESS

The apparatus used for the experiment is schematically shown in Fig. 1. A test wire is clamped between two steel vises *P*. The upper pin vise is rigidly mounted while the lower pin vise is free to rotate. *W* is a weight resting on the coupling *C*. The moment of inertia of the system is kept constant when different weights are suspended on the wire by varying the weight of the soft iron pieces I on the long inertia arm A. Consequently, the frequency of vibration does not change during the measurements. A free torsion vibration is started by two electromagnets located in front of the two iron pieces *I*, below and above the paper plane. The decay of the vibration is measured optically by means of a light beam reflected from a mirror to a scale or to a moving film camera. In order to determine the residual damping, the test wires are placed in a solenoid *S* and can be subjected to a magnetic field of about 800 oersteds. The

[1] R. Becker and M. Kornetzki, Z. Physik **88**, 634 (1934).
[2] Max Kornetzki, Wiss. Veröffentl. Siemens Werken **17**, 410 (1938).
[3] Bonfiglioli, Ferro, and Montalenti, Phys. Rev. **86**, 959 (1952).
[4] A. W. Cochardt, Trans. Am. Soc. Mech. Engrs. **75**, A196 (1953).
[5] E. R. Parker, Trans. Am. Soc. Metals **28**, 661 (1940).

brass bar B is immersed in oil so that lateral movements of the rod are damped quickly.

Measurements were made on wires 1, 2, and 3, whose compositions and dimensions are listed in Table I. The carbon content of the Puron wire (No. 3) was 0.006 percent. Alloys 1 and 3 were vacuum-melted, forged, and then swaged; alloy 2 was melted in air, forged, and centerless ground. Before the measurements were made, all wires were heat-treated at 900°C for two hours in a hydrogen atmosphere and were cooled at a rate of 120°C/hr. X-ray photographs taken on alloy 2 indicated a slight preferred orientation of the surface of the wire.

The results for wires 1 and 2 are shown in Figs. 2 and 3 for a frequency of 0.4 cycle per second. The logarithmic decrement δ, which is related to the total mechanical energy loss per cycle $\langle \Delta U \rangle_t$ and the total strain energy U_t according to the expression

$$\delta = \frac{1}{2} \frac{\langle \Delta U \rangle_t}{U_t} \qquad (1)$$

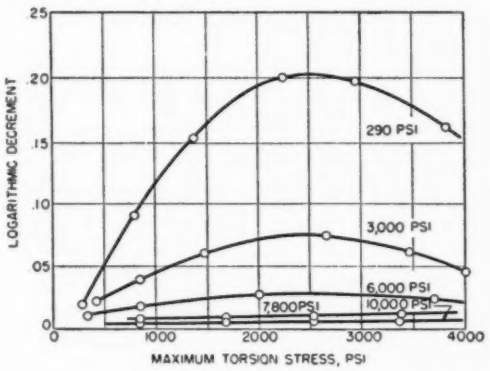


FIG. 2. Logarithmic decrement *vs* maximum torsion stress measured on an Fe-40 percent Co wire for various superimposed static tension stresses.

is plotted against the maximum torsion stress τ_m. It is evident from the figures that the static stress has a pronounced effect on the damping of the vibration. For example, at a maximum vibrational torsion stress of 2500 psi, the logarithmic decrement is 0.20 for 290 psi and 0.006 for 10 000 psi static tension stress for the Fe-40 percent Co wire (Fig. 2); in other words, the damping of the torsion vibration decreases by a factor of 33 in this range when a heavy weight is suspended on the wire. This is exemplified by Fig. 4 which shows the oscillogram for these two cases as recorded by the moving film camera. The effect of the static stress is similar for the other two wires, but not quite so strong (Figs. 3 and 5).

The appreciable difference in the damping behavior of the wires, when weights are suspended on them, is explainable in terms of domain wall movement. Under the application of the static tension stress, ferromagnetic domains become aligned along one axis, and they remain aligned if the vibrational stress is small compared

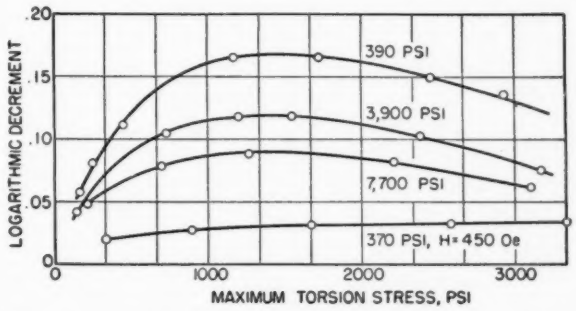


FIG. 3. Logarithmic decrement *vs* maximum torsion stress measured on an Fe-13 percent Cr wire for various superimposed static tension stresses.

to the static stress. Hence, essentially no energy is dissipated during the vibration, and the damping capacity is small.

EFFECT OF AN ALTERNATING FIELD

In a second series of measurements the effect of an alternating magnetic field on the internal damping was studied. The apparatus used for these measurements is identical with the one shown in Fig. 1 if the weight W is removed. Again, the decay of the free torsion vibration was measured. However, this time the two leads of the solenoid S were connected to an amplifier of a low-frequency oscillator. By this means not only the magnitude of the magnetic field, but also its frequency could be varied. Measurements were made on wire 4 (Table I) which was prepared and heat-treated like wire 2.

Before discussing the results, it must be mentioned that it was found more convenient for this series of measurements to express the damping capacity by the mechanical energy absorbed in the material per cycle and unit volume for a given stress. This quantity, which is given by the area of the stress-strain hysteresis loop, is denoted hereafter by ΔU. It is directly measurable only if the stress does not vary throughout the material. Since this condition is not fulfilled for a solid wire twisted in torsion, ΔU must be derived from $\langle \Delta U \rangle_t$. Both $\langle \Delta U \rangle_t$ and ΔU are considered to be functions of

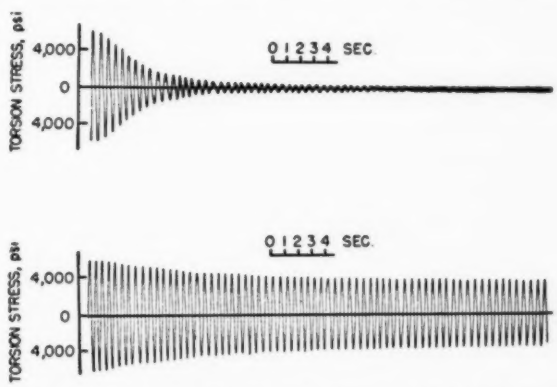


FIG. 4. The decay of a free torsion vibration of a Fe-40 percent Co wire. Above: with no weight; below: with a heavy weight suspended on the wire.

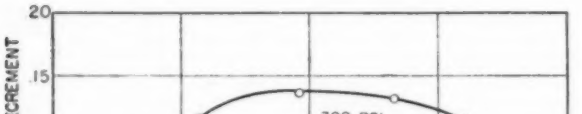


FIG. 5. Logarithmic decrement *vs* maximum torsion stress measured on a Puron wire.

the stress only; they are then related[6] to each other through the expression

$$\langle \Delta U \rangle_t = \int \Delta U dV = \int_0^{\tau_m} \Delta U \frac{dV}{d\tau} d\tau,$$

or

$$\Delta U = \frac{1}{V}\left\{ \langle \Delta U \rangle_t + \tfrac{1}{2}\tau_m \frac{d\langle \Delta U \rangle_t}{d\tau_m} \right\} \tag{2}$$

for simple torsion if $2V(\tau/\tau_m^2)$ is substituted for $dV/d\tau$. (V is the volume and τ the torsion stress.)

Figure 6 shows the values of ΔU derived in this manner as a function of the torsion stress. The magnetic field was kept constant at 6 oe while its frequency was varied. It is evident that ΔU, when plotted *versus* frequency, reaches a minimum for a frequency of the order of 20 cycles. Both at larger and at smaller frequencies the damping is higher. The increase of ΔU at higher frequencies is understandable since, due to the skin effect, the alternating magnetic field penetrates into the wire only to a certain depth.

This depth is given[7] by

$$\delta = \frac{(\rho c^2)^{\frac{1}{2}}}{2\pi (f\mu)^{\frac{1}{2}}}$$

where ρ is the resistivity, c the light velocity, f the frequency, and μ the permeability. Substituting into this expression $\rho = 10^{-5}$ ohm cm and $\mu = 10^3$, it is evident that δ is only of the order of 10^{-2} cm for a frequency of 2000 sec^{-1}; in other words, a large part of the wire is not subjected to a magnetic field any more, and this causes the increase of the damping at larger field frequencies. The increase at low frequencies, on the other hand, is more difficult to explain. Already Parker has shown that a 60-cycle magnetic field reduces the damping more than a static magnetic field of the same magnitude.[5] But he was not able to give an explanation for this phenomenon.

As was pointed out above, ΔU is related to the area of the magnetic hysteresis loop designated here by ΔU_m. The experiments showed that ΔU decreases when an alternating field is applied. Evidently, ΔU_m must then also become smaller, when measured with a superimposed alternating magnetic field, if the damping of Fe-40 percent Co wire is actually caused by magneto-mechanical hysteresis. To check this assumption magnetic measurements, similar to the mechanical measurements, were carried out. The apparatus used for determining ΔU_m is shown in Fig. 7. It is seen that the arrangement basically consists of three circuits, each connected to an Epstein frame E. The wire is replaced by Epstein strips. These have the composition Fe-35 percent Co-0.5 percent Cr and received the same heat-treatment as the Fe-35 percent Co wire. A specially built low-frequency generator S consisting of an electrolyte with two stationary and two rotating electrodes furnishes the necessary current with a constant frequency of 0.4 cycle per sec. ΔU_m is determined from the

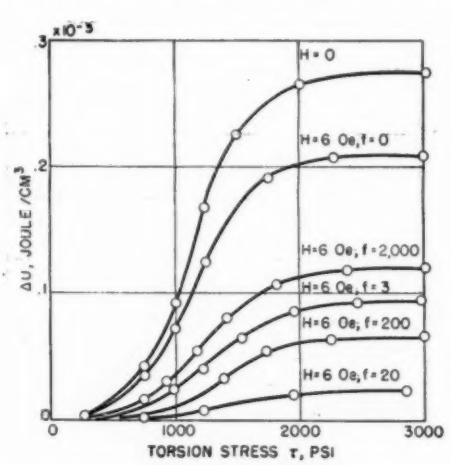


FIG. 6. Mechanical strain energy loss ΔU *vs* torsion stress τ for various frequencies f of a superimposed alternating field of 6 oersted.

reading of the wattmeter W whose current connections are in the primary circuit, while its voltage leads are in the secondary circuit. In this manner the power loss in the primary circuit due to the ohmic resistance of the primary coil does not cause any reading of the wattmeter; and since eddy current losses do not noticeably occur at the applied frequency, the deflection of the wattmeter gives only the hysteresis loss ΔU_m. B_1 and B_2 are parallel resonance circuits which prevent the alternating current, generated by the oscillator O and amplified by the amplifier P in the tertiary circuit, from entering the primary and secondary circuit. Analogously B_3 is a series resonance circuit which supresses feedback of the primary circuit into the tertiary circuit. The magnitude of the magnetic fields is determined from the reading of the ammeters A_1 and A_2. Both the wattmeter W and the Epstein frame E were described in considerable detail elsewhere.[8]

Figure 8 gives the results of the measurements, again

[6] A. W. Cochardt, Trans. Am. Soc. Mech. Engrs. (to be published).

[7] R. Becker and W. Döring, *Ferromagnetisms* (Springer-Verlag, Berlin, 1938), p. 230.

for a constant maximum alternating field of 6 oersted. The energy loss ΔU_m is plotted *versus* H_m, which is the maximum 0.4-cycle magnetic field strength. The frequency of the superimposed magnetic field was thereby varied between 0 and 2000 cycles. It is seen that the curves have the same general shape as those in Fig. 6. Again, the superimposed static field reduces the energy loss by only a slight amount. The lowest curve is the one measured at a field frequency of 20 cycles which is also in agreement with the mechanical measurements. That ΔU_m increases at higher frequencies is again the result of the skin effect. No attempts were made to correlate mechanical and magnetic energy losses since this was done several times before[9–11] and since the two materials did not coincide completely.

The described experiments explain in a simple way the effect of an alternating field on the damping of mechanical vibrations: A superimposed alternating field makes the magnetic hysteresis loop smaller; this in turn reduces the area of the stress-strain hysteresis loop, since loops are related to each other. The area of the stress-strain loop now determines the amount of damping. Hence, an alternating magnetic field must reduce the damping of a mechanical vibration.

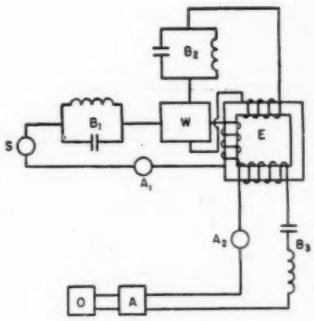


FIG. 7. Circuit for hysteresis loss measurement with superimposed alternating magnetic fields.

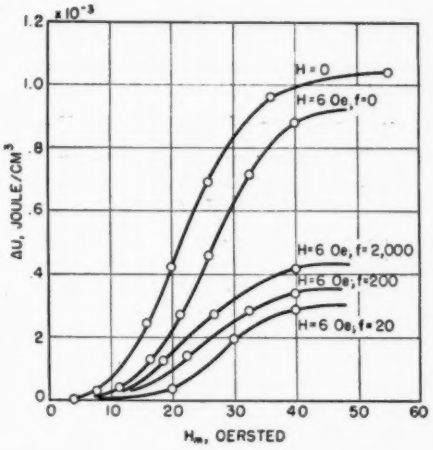


FIG. 8. Magnetic hysteresis loss ΔU_m *vs* 0.4-cycle field H_m for various frequencies f of a superimposed alternating field of 6 oersted.

SUMMARY

Ferromagnetic wires were tested in torsion, and the decay of the free vibration was measured as a function of vibrational stress, static stress, and superimposed alternating magnetic field. The effect of a static stress was studied simply by hanging different weights on the wire and by measuring the logarithmic decrement of the vibration. It was observed that, similar to a static magnetic field, a static stress decreases the damping considerably. For example, the logarithmic decrement of a Fe-40 percent Co is reduced by a factor of 33 when a heavy weight is suspended on the wire. The result of these measurements agrees with considerations of domain wall movement.

In the second part of this work, the effect of an alternating field was studied. A Fe-35 percent Co wire is subjected to an alternating magnetic field whose peak value is kept constant at 6 oersted while its frequency is varied between 0 and 2000 cycles/sec. When plotted *versus* the frequency of the field, a sharp minimum of the damping is found at about 20 cycles/sec. In order to explain this finding in terms of magneto-mechanical hysteresis, magnetic hysteresis loss measurements are made on Epstein strips of a Fe-35 percent Co, 0.5 percent Cr alloy. Again, the superimposed magnetic field is kept constant at 6 oersteds while its frequency is varied. As expected, the magnetic hysteresis loss, when plotted *versus* the frequency of the superimposed field, reaches a sharp minimum at a frequency of the order of 20 cycles/sec. Thus, the superimposed alternating magnetic field reduces the area of the stress-strain loop as well as the area of the magnetic hysteresis loop, which is in complete agreement with the magneto-mechanical effect.

ACKNOWLEDGMENTS

The author wishes to thank Dr. C. Zener, Associate Director of the Westinghouse Research Laboratories, for numerous stimulating discussions, and Mr. A. G. Hartwick for his capable assistance in taking the measurements.

[8] S. L. Burgwin, Proc. Am. Soc. Testing Materials 41, 779 (1941).
[9] M. Kornetzki, Z. Physik 121, 560 (1943).
[10] Max Kornetzki, Ann. Physik 2, 265 (1948).
[11] William Fuller Brown, Jr., Phys. Rev. 75, 147 (1949).

Letters to the Editor

Electron Bunches of Short Time Duration*

E. W. ERNST AND H. VONFOERSTER
University of Illinois, Urbana, Illinois
(Received February 4, 1954)

FOR many experimental purposes, it is necessary to have a means of generating bunches of electrons with a duration of about 10^{-12} second. This problem occurred during an investigation of the time dispersion of secondary emission. For this particular purpose, an electron bunch of this duration, recurring at regular intervals, and of a relatively small current density was required. The method which was devised to generate these bunches is not overly complex and may be of interest to experimenters faced with a similar problem.

The method of producing electron bunches of this short duration consists of two main steps. First, a group of electrons is periodically selected by a process of chopping; second, this chopped group of electrons is then compressed into a bunch of the desired duration. A diagram of the method is shown in Fig. 1.

In the chopping section, a dc electron beam is injected between the two wires of a Lecher wire pair,[1] which may be excited at a frequency of 3000 megacycles, causing the beam to be deflected across a chopping slit such that only during a definite portion of the rf period are electrons allowed to pass. To enable selection of a particular portion of the rf period, a pair of dc deflection plates is provided to direct the electron beam such that the desired portion of the trace will coincide with the chopping slit.

The group of electrons selected in this manner is then passed through the compression section where a cavity, resonant at the same frequency of 3000 megacycles, velocity modulates the group of electrons preselected by the chopper. By suitable adjustment of the relative phase of the rf energy supplied to the chopping deflector and that supplied to the compression cavity, the earlier electrons of the group will be decelerated and the later electrons will be accelerated so that all the electrons will arrive at the target plane nearly simultaneously. The quantitative aspects of this process are well understood as first-order bunching theory which is extensively treated elsewhere.[2]

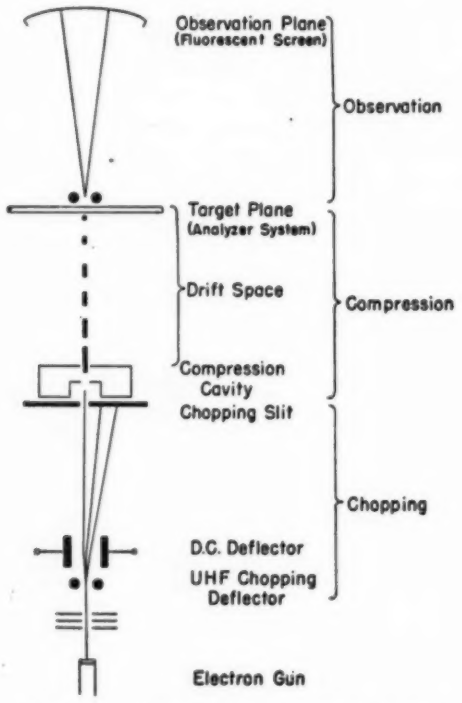


FIG. 1. Schematic diagram of a system for producing electron groups of short time duration.

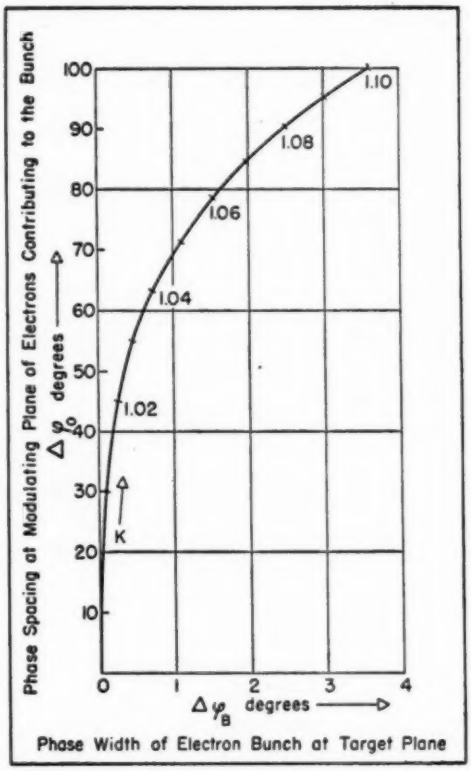
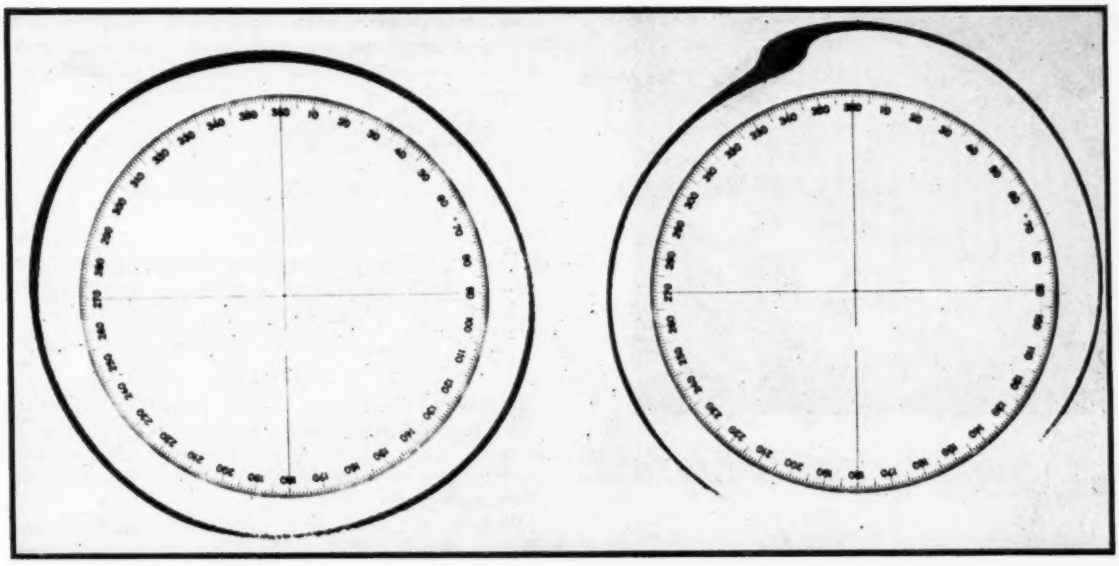


FIG. 2. Phase compression diagram for the compression section of the electron bunching system.

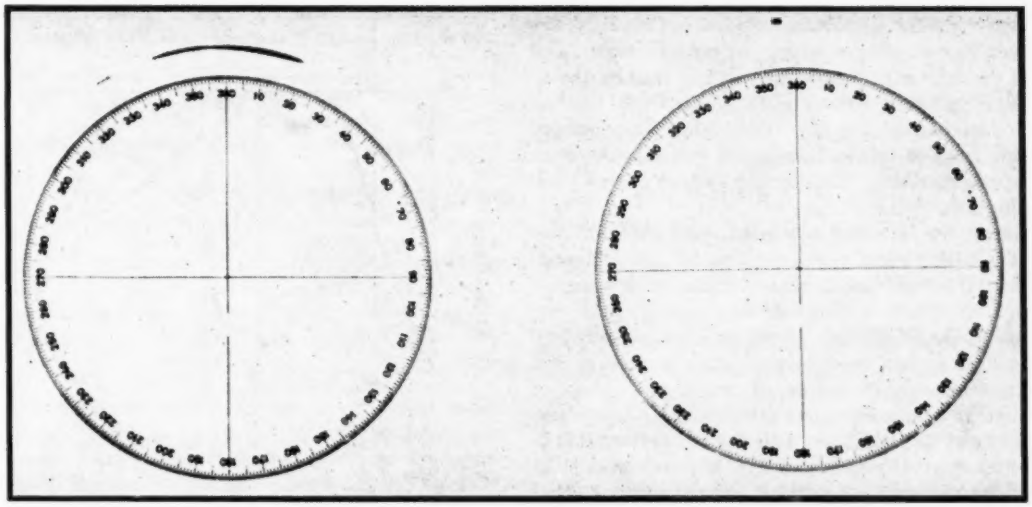
The theory indicates that, for values of the bunching parameter k, slightly greater than unity, two close bunches will be formed. Since most of the electrons are to be found at and between these two bunches, in cases when the two bunches are extremely close together this group of electrons may, for all practical purposes, be considered as a single bunch. Because it is desired to include at the target plane only electrons which contribute to this single bunch, having a phase width of $\Delta\varphi_B$, it is of interest to determine the phase spread $\Delta\varphi_0$ over which these electrons may have been distributed originally at the modulation plane. This relationship may be theoretically determined by a simultaneous solution of three transcendental equations. This solution, carried out by a numerical method, is shown in Fig. 2 where the high compression of the original phase spread $\Delta\varphi_0$ is clearly demonstrated.

In order to analyze the proper functioning of the chopping and compression sections, both separately and together, an analyzing system was placed at the target plane. This analyzer consists of two uhf deflection systems, of the same type as that used for the chopping deflector, placed perpendicular to one another and excited such that a dc beam of electrons is caused to trace a complete circle on a fluorescent screen. The action of the analyzer is to transform an electrical phase element at the target plane into a geometrical deflection angle. Thus, if a group of electrons pass the target plane during a phase interval of, say, 17°, a 17° sector on the fluorescent screen will be observed.

A set of typical traces obtained by operating the two sections, both separately and together, is shown in the photographs of Fig. 3(a) to (d). The first figure (a) shows the trace produced by



(a) (b)



(c) (d)

FIG. 3. An analysis of the current at the target plane of the electron bunching system under various conditions of operation: (a) without chopping or compression; (b) without chopping but with compression; (c) with chopping but without compression; (d) with both chopping and compression. (Original diameter of circular trace about 6 cm. Circular sweep speed about twice the velocity of light.)

a dc beam of electrons passing the target plane. Applying a small amount of velocity modulation (bunching parameter approximately 1.2), results in a concentration of electrons in the bunch region and a scarcity of electrons one-half period away, as seen in (b). The action of the chopping section alone in preselecting a group of electrons is shown in (c). Finally the operation of both the chopper and the compressor produces a result which is shown in (d). The group of electrons which occupied a phase spread of about 60° at the modulating plane has been compressed into a bunch with a phase width of approximately 2°. Since the frequency of modulation used is about 3000 megacycles, a 2° phase spread corresponds to a duration of approximately 2×10^{-12} second.

* This study is sponsored by the U. S. Air Force Cambridge Research Center.

[1] L. R. Bloom and H. M. VonFoerster, Beam Analyzer, Technical Report No. 5-3, U. S. Office of Naval Research Contract (unpublished).

[2] Karl R. Spangenberg, *Vacuum Tubes* (McGraw-Hill Book Company, Inc., New York, 1948), Chapter 17.

An Effect Produced by Oblique Impact of a Cylinder on a Thin Target

WILLIAM A. ALLEN, JOE M. MAPES, AND WESLEY G. WILSON
Michelson Laboratory, U. S. Naval Ordnance Test Station, Inyokern, China Lake, California

(Received Februrary 8, 1954)

RIGHT circular steel cylinders have been fired at conventional ordnance velocities into thin plane lead targets. If the striking angle is zero or less than a certain critical angle 2φ, the front surface of the cylinder acquires negligible plastic deformation as a result of the impact. If the critical angle 2φ is exceeded, the front surface of the cylinder is marked by a series of ridges, the spacing of which depends on the thickness of the target.

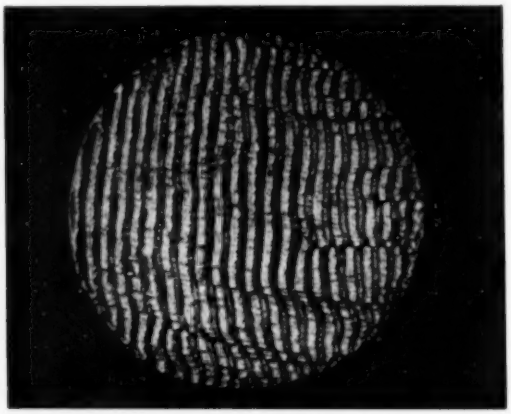


FIG. 1. End view of 0.50 caliber cylindrical steel projectile after it has passed through a 0.005-in. lead target at 45 deg striking angle. The actual diameter of this cylinder is 0.428 in.

Figure 1 is a photograph of the end of a 0.50 caliber cylindrical steel projectile after it has penetrated a 0.005-in. lead target aligned at a 45 deg striking angle. Figure 2 is a photomicrograph of a cross section of a cylinder that shows the wave structure of the ridges. Although the mechanism responsible for production of these waves is somewhat obscure, the critical angle 2φ is believed to be the same critical angle discussed by the Los Alamos group[1] in a paper that deals with metal plates accelerated together by high explosive charges. The Los Alamos group has discussed the asymmetric collision of dissimilar solids, but has not yet reported any experimental data. The specimens shown in Figs. 1 and 2 correspond to the asymmetric case.

The experiment was modified to obtain symmetric collision. Steel projectiles with conical noses specified by the half-angle $\pi/2-\theta$ were fired into steel targets aligned at the striking angle θ. Plastic deformation occurs along one of the elements of the cone provided that $2\theta>2\varphi$. Negligible plastic deformation occurs if $2\theta<2\varphi$. The critical angle 2φ determined by this experiment is in excellent agreement with the predicted[1] value for iron. Two preliminary determinations indicate the value $2\varphi=7.7$ deg for a projectile velocity of 0.87 mm/μsec. This velocity corresponds to the plate velocity $v_p=0.43$ mm/μsec of Fig. 15 in reference 1.

The experiment discussed is believed to be equivalent to that of the Los Alamos group. No theoretical or experimental difficulty is expected if the technique is extended to higher velocities and to solids other than steel. The experiment is expected to be of value in checking and determining equation of state data of solids in the megabar pressure regime. As a basis for comparison, the compressibility of pure iron has been measured up to a maximum pressure of only 0.03 megabar.[2] The pressure produced at the critical angle 2φ by the symmetric collision of steel is 0.47 megabar, a value calculated from the published[1] equation of state of iron.

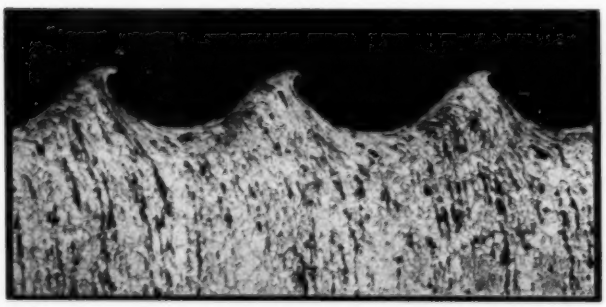


FIG. 2. Photomicrograph of a cross section of a projectile which shows the wave structure formed by 45 deg impact on a 0.010-in. lead target. The average distance from crest to crest is 0.033 in.

[1] Walsh, Shreffler, and Willig, J. Appl. Phys. **24**, 349 (1953).
[2] P. W. Bridgman, Revs. Modern Phys. **18**, 1 (1946).

A New Silicon *p-n* Junction Photocell for Converting Solar Radiation into Electrical Power

D. M. CHAPIN, C. S. FULLER, AND G. L. PEARSON
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received January 11, 1954)

THE direct conversion of solar radiation into electrical power by means of a photocell appears more promising as a result of recent work on silicon *p-n* junctions. Because the radiant energy is used without first being converted to heat, the theoretical efficiency is high.

Photons of 1.02 electron volts ($\lambda=1.2$ microns) are able to produce electron-hole pairs in silicon. In the presence of a *p-n* barrier, these electron-hole pairs are separated and made to do work in an external circuit. All of the light of wavelength shorter than 1.2 microns is potentially useful for generating electron-hole pairs but the efficiency of energy conversion decreases for short wavelengths because the energy above the necessary 1.02 electron volts is wasted. Allowing for this loss and assuming a working voltage of 0.5 volt, which is near the maximum measured, a computation over the entire solar spectrum indicates a limiting efficiency of approximately 22 percent for a cell of negligible internal losses and for utilization of all possible electron-hole pairs.

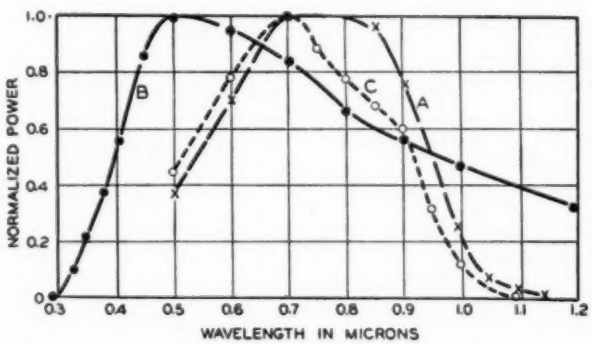


FIG. 1. Normalized spectral energy distribution. (A) Silicon photocell equi-energy response. (B) Solar energy at earth's surface. (C) Curve *A* times Curve *B*.

Several practical factors lower this figure. The untreated silicon surface reflects about half of the incident radiation. Some of this can be saved by proper surface treatment. The second serious loss is recombination of electron-hole pairs before they reach the *p-n* barrier. Penetration of radiation over most of the useful spectrum is extremely shallow so that it becomes necessary to place the *p-n* junction as near to the surface as possible except for the third serious loss. This is the I^2R loss caused by resistance in the surface layer and by contact resistance. Extremely small cells minimize the resistance loss and give useful data. For cells of several square centimeters, special geometry of contacts will minimize resistance losses.

Present work on silicon *p-n* photocells uses a thin layer of *p*-type silicon formed over an *n*-type base. The surface layer is less than 0.0001 inch thick. Figure 1 shows the spectral response for one such cell. Curve *A* is the measured power output for equal intensities of weak radiation as a function of wavelength. Maxi-

mum sensitivity is arbitrarily taken as unity. Ideally the maximum should appear near 1.2 microns with half this sensitivity at 0.6 and one-quarter at 0.3 micron. Curve *B* is the normal distribution of full sunlight[1] with the maximum expressed as unity. Curve *C* is the product of Curve *A* and Curve *B* again reduced to unity for maximum. This curve shows which part of the sun's radiation is most useful for this particular cell.

The photocells described here have been made to deliver power from the sun into a resistance load at the rate of 60 watts per square meter of photocell surface. This is approximately 6 percent efficiency and compares with a measured value of 0.5 percent on a commercially available photocell. The greatest over-all efficiency previously reported for direct conversion of solar radiation into electrical power is that of Telkes[2] using thermoelectric junctions and amounts to 1 percent.

We wish to thank H. B. Briggs who made the spectral measurements shown in Curve *A*.

[1] W. E. Forsythe, *Measurement of Radiant Energy* (McGraw-Hill Book Company, Inc., New York, 1937).
[2] Maria Telkes, J. Appl. Phys. **18**, 1116 (1947).

Dissected Amplifiers Using Negative Resistance

W. SHOCKLEY AND W. P. MASON
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received December 28, 1953)

IT is the purpose of this communication to point out a line of attack upon the problem of making high-frequency semiconductor amplifiers. This attack may be described as the dissection of the amplification mechanism into its two constituents: negative resistance and directionality.

Negative resistance is a common feature of all amplifying devices operating from a dc power source. Thus the net ac power from the device when operating as an amplifier is positive. Since this power is the net flux of the ac part of Poynting's vector, it follows that Poynting's vector must have a positive divergence throughout certain "negative resistance regions." For a vacuum-tube triode amplifier, the grid-plate space is such a region, the current being high when the voltage is low and vice versa. Similarly, the space-charge region of the collector junction of a junction transistor amplifier has negative resistance. It should be noted that these negative resistances are not characteristic of the regions *per se* but instead of their behavior in an amplifier circuit.

This method of analysis enables one to evaluate quickly such proposals as the making of a junction transistor from isolated *p-n* junctions (impossible because no negative resistances can occur) or an amplifier from a nonlinear dielectric and dc power source.

Directionality arises in vacuum tubes and transistors because of the separation of input and output circuits. This leads to asymmetry in the current voltage matrix and permits the achievement of high gain by cascading stages without resultant instability.

Because the drift velocity of electrons (or holes) in semiconductors is smaller than the average transit velocity in vacuum tubes, the physical dimensions must be smaller for the same transit times. Although the solid nature of transistors gives them an inherent structural advantage over vacuum tubes, the smaller scale required for them poses serious problems of construction for devices in the highest-frequency ranges.

One way around this difficulty is to make two-terminal negative resistance devices in which only one dimension need be small. A variety of operating principles are possible for the operation of such devices.[1]

Two-terminal negative resistance devices have the disadvantage, however, that they do not have directionality; consequently, high gain is associated with narrow margins of stability.

This limitation may be overcome by adding directionality through passive elements[2] such as Hall effect couplers or gyrators[3]

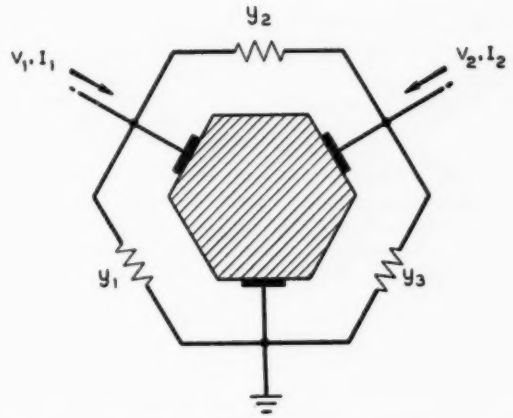


FIG. 1. Hall effect plate and three resistances. (Power supplies and connections necessary to power the negative resistances are not shown.)

in wave guides.[4] Such combinations can simulate conventional amplifying devices such as a vacuum-tube triode. In Fig. 1 the hexagon represents a germanium plate with a magnetic field perpendicular to its plane. It contributes in Eqs. (1) and (2) the terms involving the conductance y_0 and the dimensionless quantity α which is an odd function of the Hall angle with $|\alpha|<1$:

$$I_1=(y_0+y_1+y_3)V_1-[\tfrac{1}{2}(1-\alpha)y_0+y_2]V_2, \quad (1)$$

$$I_2=-[\tfrac{1}{2}(1+\alpha)y_0+y_2]V_1+(y_0+y_2+y_3)V_2. \quad (2)$$

The conductances y_1, y_2, and y_3 may correspond to negative resistance elements. For example, if y_1, y_2, and y_3 have the values

$$y_1=y_3=-(1+\alpha)y_0/2, \quad (3)$$

$$y_2=(\alpha-1)y_0/2, \quad (4)$$

then Eqs. (1) and (2) reduce to

$$I_1=0 \text{ and } I_2=\alpha y_0 V_1. \quad (5)$$

This corresponds to an ideal pentode for $\alpha>0$ and to such a pentode with voltage reversing transformer for $\alpha<0$. Thus the circuit can simulate conventional amplifiers and in addition produce new types of devices.

Values of α as high as 0.2 can be obtained in *n*-type germanium at room temperature at fields of 17 500 oersteds. Similar reasoning to that discussed here applies also to the four-terminal Hall effect plate or balanced gyrator.[3]

A simple test[5] of this principle was made at low frequencies by putting two gas diodes with negative resistance characteristics on either side of the balanced gyrator of reference 3. With this arrangement a gain of 6 db was obtained in one direction, and a loss of 46 db in the other. By more careful balancing of the negative resistances with the input and output resistances of the gyrator, higher gains are possible without instability.

[1] A guide to one method of designing such semiconductor devices is furnished by the work of F. B. Llewellyn on transit time vacuum-tube diodes: Bell System Tech. J. **13**, 59–101 (1934); **15**, 575–586 (1936). Methods of extending this approach to semiconductors have been indicated by W. Schockley, U. S. Patent 2,623,102 (Junction Transistor). Another physical effect considered by one of us (W.S.) in unpublished work is the negative differential mobility to be expected for holes in very high electric fields. Interest in this predicted effect was the provoking cause of the research on high field mobilities, see E. J. Ryder and W. Shockley, Phys. Rev. **81**, 139 (1951) and E. J. Ryder, Phys. Rev. **90**, 766–769 (1953). The theory of this effect has been independently discovered by H. Krömer, Z. Physik **134**, 435 (1953), who calls it the "Staueffect" but does not consider its utilization as a power source.
[2] Passive elements can have the desired directionality, i.e., lead to unsymmetrical current-voltage matrices, only in the presence of magnetic fields. For a proof see H. B. G. Casimir, Revs. Modern Phys. **17**, 343–350 (1945).
[3] Mason, Hewitt, and Wick, J. Appl. Phys. **24**, 166–175 (1953). The fact that transmission through a crystal is nonreciprocal in the presence of a magnetic field was first pointed out by E. M. McMillan, J. Acoust. Soc. Am. **19**, 922 (1947).
[4] C. L. Hogan, Bell System Tech. J. **31**, 1–32 (1952).
[5] This test was made by W. H. Hewitt.

On the Interaction of Two Plane Shocks Facing in the Same Direction

OTTO LAPORTE AND E. B. TURNER
Randall Laboratory of Physics, University of Michigan, Ann Arbor, Michigan
(Received January 4, 1954)

THE problem of the coalescence of two shocks has been given considerable theoretical attention in the literature. First treated by von Neumann[1] and by Courant and Friedrichs,[2] this case of shock interaction was calculated in considerable detail by Bull, Fowell, and Henshaw.[3] The time sequence of events is as follows. First there must be two shocks separated by a region of constant flow; at the moment when the rear shock reaches the front shock, as it always must, there is formed a stronger shock of coalescence proceeding in the same direction as well as an interface which moves with the flow and a centered rarefaction which moves backwards. For γ larger than 5/3, a case which will not be considered further here, the rarefaction may be replaced by a weak shock facing in the direction opposite to the amalgamated shock.

As far as we know this phenomenon has up to now escaped observation in shock tubes. We should like to report a method which makes it possible to create two equally facing shocks separated by a constant state. This is accomplished by blocking the test section with a comblike stack of plates spanning the width of the tube (see Figs. 1 and 2). Evidently the wedgelike tips will, when struck by the incident shock, give rise to cylindrical reflected shocks which quickly, by Huygens principle and Mach interaction, form the first plane shock. In the meantime segments of the primary shock travel down into the interstices, are reflected at the ends, and on emergence form a second plane shock. The strengths of these two similarly facing shocks may evidently be varied between wide limits by (a) varying the strength of the primary shock and (b) varying the ratio of the thickness of the plates to that of the interstices.

The two interferograms shown below were taken with the University of Michigan 2- by 7-inch shock tube, using the 9-inch, 30°–60° Mach-Zehnder interferometer. The gas used was argon, and the shock strength or pressure ratio of the incident shock was 12. Seven plates separated by six spacers of $\frac{3}{4}$-inch thickness were employed. The first interferogram, taken with white light, shows two perfectly formed shocks[4] proceeding away from the stack of plates. The fringe shifts indicate that they possess approximately the same density discontinuity and that they are

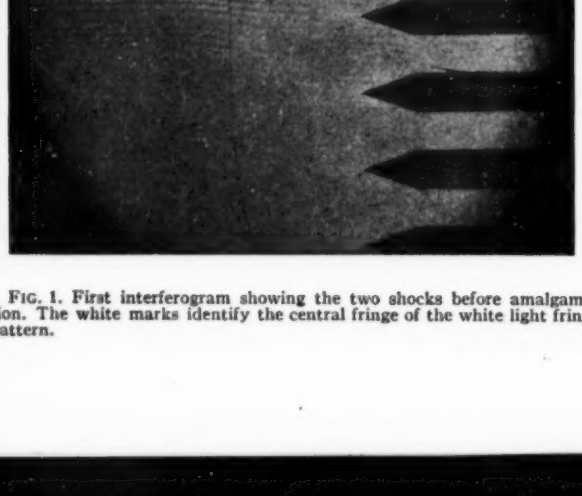


FIG. 1. First interferogram showing the two shocks before amalgamation. The white marks identify the central fringe of the white light fringe pattern.

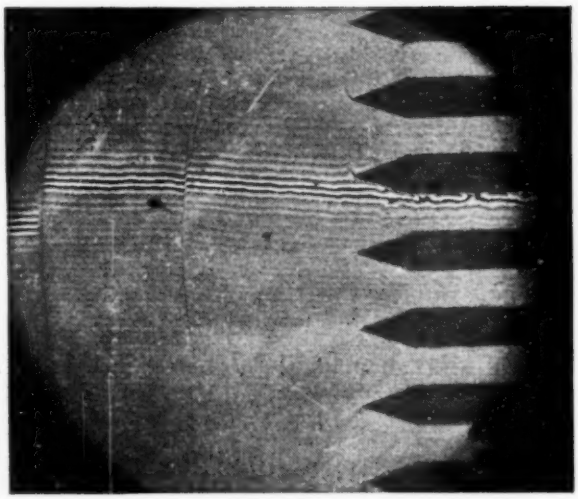


FIG. 2. Second interferogram taken after interaction.

separated by a region of constant flow. The second interferogram shows at the extreme left the amalgamated shock and its larger fringe shift. Somewhat to the right the interface with its minute fringe shift may be seen. According to the theory, a rarefaction should be present between the interface and the reflecting device, but it is not discernible in our flow pictures. The reason for this is twofold: First, for the shock strengths used here, the density change across this rarefaction is small compared with the density change across the interface (which itself is small in comparison to the density change across the resulting shock). Secondly, inasmuch as plate-glass windows were used on the shock tube, it was not intended to observe the phenomenon in its ultimate refinement at this point. The density changes inferred from the fringe shifts are in accord with the theory within our experimental error. Schlieren photographs, taken of the same process, show the interface with extraordinary sharpness and clarity.

This work was partially supported by the U. S. Office of Naval Research under Contract No. N6-onr-232, T. O. 4, with the Engineering Research Institute of the University of Michigan.

[1] J. von Neumann, Progress Report on the Theory of Shock Waves. National Defense Research Committee, Division 8, Office of Scientific Research and Development No. 1140, 1943.
[2] R. Courant and K. O. Friedrichs, Interaction of Shock and Rarefaction Waves in One-Dimensional Motion. National Defense Research Committee, Applied Mathematics Panel Report 38.1R (Applied Mathematics Group—New York University No. 1), 1943.
[3] Bull, Fowell, and Henshaw, The Interaction of Two Similarly Facing Shock Waves. Institute of Aerophysics, University of Toronto, UTIA Report No. 25, 1953.
[4] Since the plate stack did not span the height of the tube entirely, but left slightly more than 1 inch at top and bottom, the shocks show a slight curvature, especially at their ends and at the larger distances from the reflecting plates.

High-Speed Magnetic Pulsing of Ferrites

R. CONWAY LECRAW
Diamond Ordnance Fuze Laboratories, Washington, D. C.
(Received December 18, 1953)

A NUMBER of papers have reported Faraday rotations of microwaves using ferrites in wave guides.[1,2] Work has been done by Bryan[3] at The Naval Research Laboratory to determine the limiting speeds of rotation for a given ferrite in connection with high-speed microwave switches. He found that, for a particular NiZn ferrite and method of pulsing, the minimum switching cycle is 0.8 μsec. This limitation was attributed primarily to the recovery time of the ferrite.

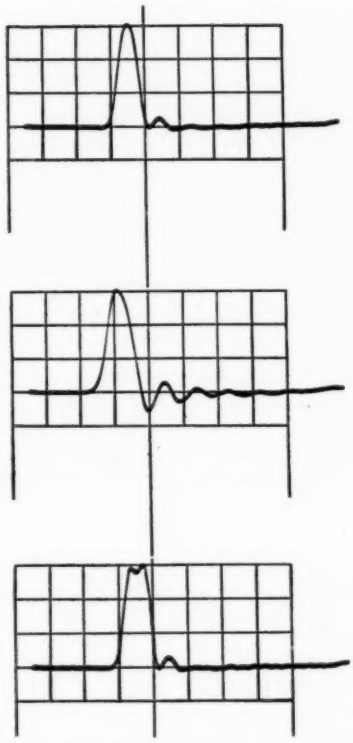


FIG. 1. (top). Video output of detecting crystal, max rotation =90°. The 1.5-db point =one-half the first vertical division.
FIG. 2. (middle). Current I in the coil, max =12 amp.
FIG. 3. (bottom). Video output of detecting crystal, max rotation >90°. The same external triggering pulse was used for all three figures. The time scale =0.025 μsec per division.

Experiments on magnetic pulsing of ferrites at this laboratory have resulted in a switch with a recovery time of 0.012 μsec to the 1.5-db point and a switching cycle of 0.025 μsec between 1.5-db points. The insertion loss is less than 1 db, and the attenuation in the switch-off state is greater than 35 db.

A rod of a MgMn ferrite is placed along the axis of a section of circular wave guide propagating a TE_{11} wave at X band. A longitudinal magnetic field produces a rotation of a linearly polarized wave due to the difference in velocities of the two oppositely rotating circularly polarized waves of which the linear wave is composed.[4] Rotations of 0° and 90° correspond, respectively, to maximum and minimum transmission into a section of rectangular wave guide terminated with the detecting crystal. A piece of resistance card just preceding this section and parallel to its long dimension absorbs the microwave energy in the switch-off state.

An eight-turn coil is wound about the circular wave guide containing the ferrite. A thin axial slit in the guide wall prevents the wall from acting as a shorted turn to the coil. The current I in the coil is produced by a hydrogen thyratron pulser and is monitored by a synchroscope across a 3-ohm coaxial-type resistor in series with the coil. This type resistor has negligible inductive reactance up to several hundred Mc. The current pulse of Fig. 2 produces the video output pulse of Fig. 1 when the maximum amplitude of I is adjusted to produce 90° rotation. The peak magnetizing field is about 18 oersteds. In Fig. 3 a larger amplitude of I produces more than 90° rotation. Some ringing is noted in the current pulse. This can be eliminated if desired. However in these experiments the recovery time of the ferrite switch is of principal interest.

An interesting and somewhat unexpected result is that the ferrite magnetization (to the first approximation proportional to rotation) follows the magnetizing current quite closely. No appreciable time delay in the magnetization process is evident. The time delay of the entire output pulse with respect to the current pulse is due to the length of wave guide between the switch and the detecting crystal. This was checked by using various lengths of wave guide.

Experiments are in progress to study the response of the magnetization of various ferrites as a function of pulse amplitude and length with the latter varying from 0.01 to 0.5 μsec.

[1] N. G. Sakiotis and H. N. Chait, Proc. Inst. Radio Engrs. 41, No. 1 (1953).
[2] C. L. Hogan, Bell System Tech. J. XXXI, No. 1 (1952).
[3] J. W. Bryan (private communication).
[4] R. C. LeCraw, Natl. Bur. Standards Rpt. No. 2313, March, 1953.

Crossed Electron Beam Technique for Measuring Space-Charge Effects in Beams*

PAUL D. COLEMAN, MURRAY D. SIRKIS, AND IRVING KAUFMAN
Electrical Engineering Department, University of Illinois, Urbana, Illinois

(Received December 28, 1953)

SPACE-CHARGE problems associated with the production and bunching of high current density electron beams occupy the attention of many workers in the electronics field. The crossed electron beam technique[1] shows promise as a method of studying space-charge effects in beams in greater detail and with more additional information than, for example, the iris technique of Linder and Hernquist.[2]

The method, as illustrated in Fig. 1, is to probe the high current density test beam by a low voltage swept probe beam and observe the Coulomb scattering on a fluorescent screen. Under the conditions of our experiment, it can be shown that the semiscattering angle θ_0 is given by

$$\theta_0 \simeq \frac{\pi q \sigma(z)}{16 \epsilon m_0 v_0^2}, \tag{1}$$

where $\sigma(z)$ is the charge per unit length of the test beam and v_0, the velocity of the probe beam at the point of crossing. Thus the deflection is proportional to the net charge per unit length, i.e., electron density minus positive ion density, and hence can be used to evaluate directly the positive ion density as a function of time.

The rate of formation of ions per cm³ in an electron beam is given by the expression

$$\frac{dn_2}{dt} = pP(V) n_1 v_1, \tag{2}$$

where p is the gas pressure, n_1 the electron density, v_1 the electron velocity, and $P(V)$ the number of ion pairs formed by an electron per unit length at unit pressure[3] (a function of voltage V).

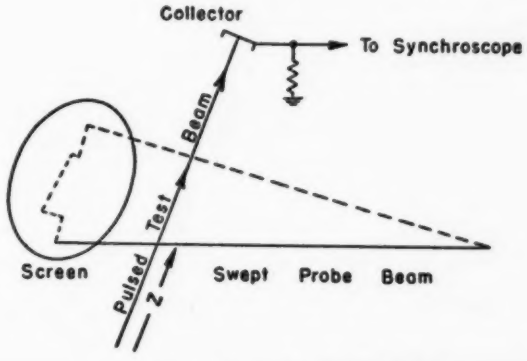


FIG. 1. Crossed electron beam arrangement.

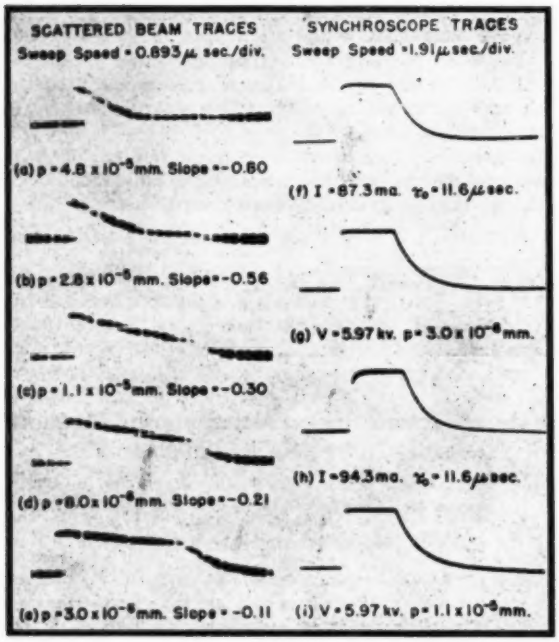

SCATTERED BEAM TRACES
Sweep Speed = 0.893 μ sec./div.
SYNCHROSCOPE TRACES
Sweep Speed = 1.91 μ sec./div.
(a) p = 4.8 x 10⁻⁵ mm. Slope = -0.60
(f) I = 87.3 ma. τ₀ = 11.6 μ sec.
(b) p = 2.8 x 10⁻⁵ mm. Slope = -0.56
(g) V = 5.97 kv. p = 3.0 x 10⁻⁶ mm.
(c) p = 1.1 x 10⁻⁵ mm. Slope = -0.30
(h) I = 94.3 ma. τ₀ = 11.6 μ sec.
(d) p = 8.0 x 10⁻⁶ mm. Slope = -0.21
(e) p = 3.0 x 10⁻⁶ mm. Slope = -0.11
(i) V = 5.97 kv. p = 1.1 x 10⁻⁵ mm.

FIG. 2. Effect of pressure on scattered probe beam for 5.97-kv electrons.

Figure 2 shows the effect of pressure on the scattered probe beam resulting from a high current density pulsed test beam and the actual current and voltage pulses measured on a synchroscope. The product n_1v_1 is readily calculated from the synchroscope traces as follows:

$$n_1v_1=\frac{I}{qA} \quad \text{with} \quad v_1=5.93\times10^7\sqrt{V}\ \text{cm/sec}, \tag{3}$$

where A (0.32 cm²) is the mean cross-sectional area of the test beam. The derivative dn_2/dt is determined from the initial slope m of the scattered trace as follows:

$$\frac{dn_2}{dt}=\frac{dx}{dt}\frac{dn_2}{dx}=-S\frac{\Delta n_2}{\Delta y/m}=-sm\frac{n_1}{\Delta y}\text{ions/cm}^3\text{/sec}, \tag{4}$$

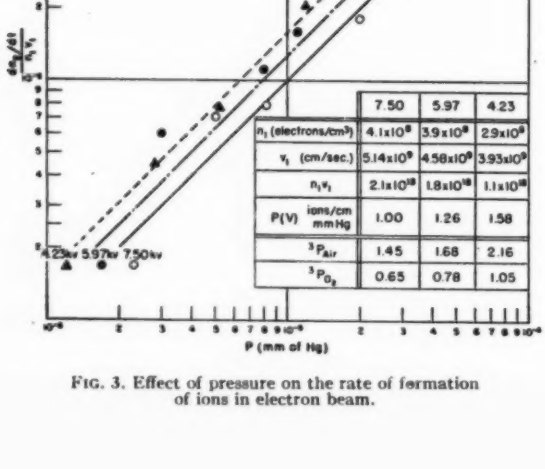


FIG. 3. Effect of pressure on the rate of formation of ions in electron beam.

where Δy = divisions initial vertical displacement and s = divisions/second.

A plot of the data taken from photographs, such as those of Fig. 2, is given in Fig. 3. A value of $P(V)$ computed from the curves is also given for three different voltages.

The scattered beam traces indicate that the test beam never becomes completely neutralized by the positive ions. The final steady-state condition will be a function of the geometry of the tube and any ion sweeping fields that may be present.

* Work supported in part by Air Force Cambridge Research Center on Contract AF 19(604)-524.

[1] P. D. Coleman and M. D. Sirkis, J. Appl. Phys. **21**, 952 (1953).
[2] E. G. Linder and K. G. Hernquist, J. Appl. Phys. **21**, 1088 (1950).
[3] Engel and Steenbeck, *Elektrische Gasentladungen* (Berlin, 1932). [$P(V)$ for O_2 extrapolated from given curves.]

A Combined Slow-Wave Circuit and Beam Focusing Structure

PHILIP RICE
Stanford Research Institute, Stanford, California
(Received January 7, 1954)

THE "strong focusing effect"[1] has been shown to apply equally well to low-velocity electron beams in the presence of space charge.[2] Such beams have been focused in periodic axially symmetrical magnetic fields with considerable savings in weight of the permanent magnetic structure. The field was produced by means of a series of ring magnets.

A familiar slow-wave circuit used in traveling wave oscillators and amplifiers consists of a series of interlocking "fingers" protruding from opposite walls of a wave guide. An axial hole permits the electron stream to travel down the device. Such a structure can be made to produce a periodic magnetic field which will focus the electron stream.

In order to test the idea, a short section consisting of five fingers was built. Figure 1 shows the circuit with the beam tester in place. All the fingers except for the first are 1.75 in. long, 1.5 in. wide, 0.25 in. thick, and 0.25 in. apart. To obtain correct

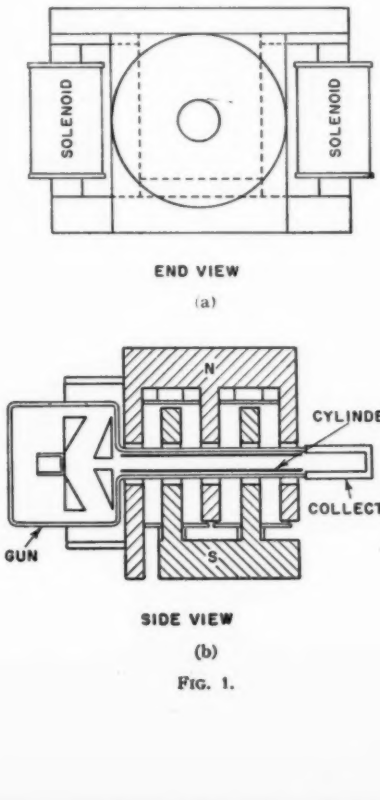


FIG. 1.

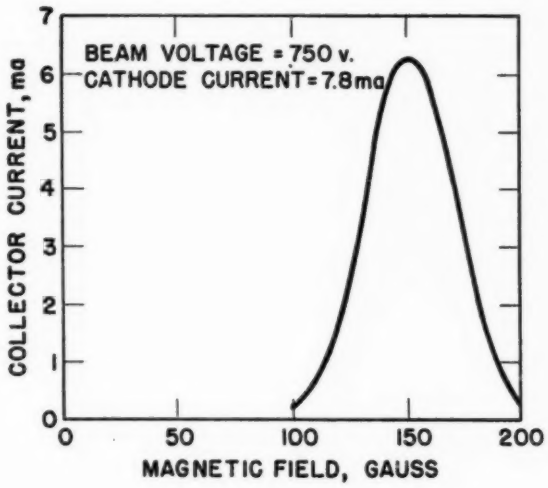


FIG. 2.

magnetic field starting conditions, the first finger was made 2.5 in. long and 2.25 in. wide and a steel cylinder 0.75 in. long was attached to it. The hole is 0.5 in. in diameter. A Pierce type electron gun which produces 14 ma at 1000 v directs a beam having a minimum diameter of 0.07 in. down a copper cylinder 2.25 in. long and 0.18 in. in diameter. The magnetic field between the fingers is supplied by four iron-cored solenoids along the edges of the structure. The field could equally well be supplied by two U-shaped permanent magnets running the length of the circuit. The entire device is made from mild steel.

Figure 2 shows collector current as a function of peak magnetic field in the first gap. The "band pass" type of transmission is evident. Higher field bands were not obtained because of saturation of the iron. The Brillouin value of field for a 0.10-in. diameter beam at 750 v and 7.8 ma is 110 gauss.

With appropriate rf shielding, this type of magnetic circuit can be used to focus helix-type tubes without the limitation on permanent magnet length inherent in a structure of ring-shaped magnets and pole pieces.

[1] Livingston, Courant, and Snyder, Phys. Rev. **88**, 1190–1196 (1951).
[2] J. R. Pierce, J. Appl. Phys. **24**, 1247 (1953).

JOURNAL OF APPLIED PHYSICS VOLUME 25, NUMBER 5 MAY, 1954

Announcements

American Physical Society—Members of the Executive Committee of the Division of Fluid Dynamics for 1954

Chairman: François N. Frenkiel, Applied Physics Laboratory, The Johns Hopkins University.

Vice-Chairman: Otto Laporte, University of Michigan.

Secretary-Treasurer: Joseph O. Hirschfelder, University of Wisconsin.

Members: Walker Bleakney, Princeton University. S. Chandrasekhar, Yerkes Observatory, University of Chicago. Bernard Lewis, Combustion and Explosive Research, Inc., Pittsburgh, Pennsylvania. Hans W. Liepmann, California Institute of Technology.

High-Vacuum Symposium

The Committee on Vacuum Techniques announces that a High-Vacuum Symposium will be held at the Berkeley Carteret Hotel at Asbury Park, New Jersey, on June 16, 17, and 18, 1954.

Papers are being selected for presentation in five major categories, which include: 1. Panel on nomenclature and standards; 2. new equipment and instruments; 3. fundamental developments in vacuum technology; 4. methods and techniques; and 5. applications and processes. Dr. K. C. D. Hickman, will be Honorary Chairman of the Symposium.

The Committee on Vacuum Techniques has been organized as a nonprofit corporation. Its prime purpose is to sponsor the high-vacuum symposium. Other aims are to further the dissemination of knowledge pertaining to high-vacuum techniques and to assist in the establishment of universally acceptable nomenclature and standards.

All individuals and companies interested in attending the Symposium, presenting papers, or participating in other activities of the Committee on Vacuum Techniques, are urged to write to P. O. Box 1282, Boston, Massachusetts.

International Colloquium on Oxide Cathodes

Fifty years ago the first paper dealing with oxide coated cathodes was published. To celebrate the fiftieth anniversary, the Société Française des Ingénieurs Techniciens du Vide is organizing an International Convention to be held in Paris on June 24 and 25, 1954. Every prominent specialist is expected to attend this convention.

Papers presented within the scope of the program may consist of original material or re-edited drafts of valuable documents. Papers accepted by the Organizing Committee will be presented by the authors or a summary will be read by the Secretary at the sessions of the convention. It is requested that these papers be sent to us, in their definitive form, no later than May 15, 1954. The texts will be published in a special issue of the periodical *Le Vide*, the official bulletin of the S.F.I.T.V. in their original language with, if possible, a French translation. Each author will receive, free of charge, a full copy of the Convention Proceedings, with 20 reprints of his own contribution.

For information, please apply to the Société Française des Ingénieurs Techniciens du Vide, 44, rue de Rennes, Paris VI°, France.

Summary of Program: I—Basic metal; II—carbonates and coating process; III—emission theory of oxide coated cathodes; IV—definition of the cathode properties; V—special cathodes; and VI—stability of emissive properties.

International Conference on Semiconductors

The Netherlands Physical Society, with the support of the International Union of Pure and Applied Physics and U.N.E.S.C.O. will organize an International Conference on Semiconductors, to be held at Amsterdam from June 29th–July 3rd, 1954.

The following scientists (two of them with some reserve) will deliver lectures: J. Bardeen, W. H. Brattain, H. B. G. Casimir, F. A. Kröger, D. Polder, M. Schön, W. Shockley, R. A. Smith,

H. J. Vink, H. Welker on subjects such as bulk recombination; surface conductivity; surface trapping; surface recombination; intermetallic compounds, the band picture in polar and nonpolar semiconductors; photoconductivity in semiconductors like PbS, PbTe, PbSe, ZnS, CdS; the application of general physical and chemical laws for the preparation of semiconductors with specific properties.

Discussions will be held in connection with these main lectures and there will be opportunity to discuss problems in more detail in sectional meetings. In these sectional meetings short communications of about 15 minutes can be given.

Scientists who would like to participate in the conference or want to give a scientific contribution should communicate this to the Secretary (Dr. H. J. Vink, Floralaan 142, Eindhoven/Holland) as soon as possible. They will receive a detailed program in due course.

Conference on Defects in Crystalline Solids

The H. H. Wills Physical Laboratory of the University of Bristol, in cooperation with the International Union of Pure and Applied Physics (particularly its Commission for the Physics of the Solid State) and with The Institute of Physics, is organizing a conference on Defects in Crystalline Solids to be held from July 13–July 17, 1954, in Bristol. While not excluding other subjects in the field, the organizers propose to give particular attention to defects such as dissolved atoms, vacancies, and *F* centers, to microwave resonance methods of investigating their properties, and to the way in which they react with dislocations. Thus dislocations will be discussed in their chemical aspects, as influencing diffusion and precipitation in the solid state, rather than in relation to plastic flow.

It is hoped that a number of authors from overseas will personally present their papers, and with this in mind the conference has been arranged to follow immediately after the General Assembly of the International Union of Pure and Applied Physics.

Board and lodging will be provided in Wills Hall, a student hall of residence, on special terms, or at hotels.

The conference is open to any scientist interested in this field, subject to the limitations of seating accommodation.

Further particulars may be obtained from the Secretary, H. H. Wills Physical Laboratory, Royal Fort, Bristol 8, or from the Secretary, The Institute of Physics, 47, Belgrave Square, London S.W.1.

Those wishing to attend the conference are asked to apply to the former, marking the envelope "1954 Conference," stating whether they wish to be accommodated at Wills Hall or at a hotel and for what nights accommodation is required.

International Union of Crystallography

As already announced, the Third General Assembly and International Congress of the Union will be held in Paris from July 21 to July 28, 1954. Full details of the meeting are now available in the First Circular, copies of which may be obtained from the General Secretary of the Union (R. C. Evans, Crystallographic Laboratory, Cavendish Laboratory, Cambridge, England), from the Secretary of the Programme Committee (A. J. Rose, Laboratoire de Minéralogie, 1 rue Victor Cousin, Paris 5, France) or from any of the following regional representatives of the Programme Committee: G. Menzer (Universitätsinstitut für Kristallographie und Mineralogie, Luisenstrasse 37/II, München 2, Germany), W. H. Taylor (Crystallographic Laboratory, Cavendish Laboratory, Cambridge, England), J. D. H. Donnay (The John Hopkins University, Baltimore, Maryland, U.S.A.).

At the congress papers will be presented on all aspects of crystallographic research; there will also be an exhibition of crystallographic apparatus and books. After the congress two specialized symposia will be held on "The Location and Function of Hydrogen" and "The Mechanism of Phase Transitions in Crystals," and there will be visits to localities of mineralogical interest. Offers of papers for the congress and symposia and notice of enrollment must reach the Secretary of the Programme Committee (preferably on the form accompanying the First Circular) by February 15, 1954. All general correspondence should also be addressed to the Secretary of the Programme Committee.

See September issue of *Acta Crystallographica* for full details.

Cooperative Graduate Summer Sessions in Statistics

Beginning in 1954, North Carolina State College, the University of Florida, Virginia Polytechnic Institute, and the Southern Regional Education Board will jointly sponsor cooperative Graduate Summer Sessions in Statistics. The first session will be conducted by a distinguished faculty at Virginia Polytechnic Institute in the summer of 1954 and will run from June 9 through July 17.

Inquiries should be addressed to Boyd Harshbarger, Head, Department of Statistics, Virginia Polytechnic Institute, Blacksburg, Virginia.

Short Courses on High-Temperature Properties of Materials and Mechanics of Creep

A short course on High-Temperature Properties of Materials will be held at The Pennsylvania State University from June 21 to June 25 inclusive. Following this course one, on Mechanics of Creep will be conducted between June 28 and July 2. Authorities from industrial, governmental, and educational institutions will give the lectures for both these courses. The programs are planned to give a broad coverage of high-temperature properties of materials and mechanics of creep by including physical, metallurgical, and design aspects of these subjects. For further information, please write to Dr. Joseph Marin, Department of Engineering Mechanics, The Pennsylvania State University, State College, Pennsylvania.

Course in Modern Industrial Spectroscopy at Boston College

Boston College has announced a special two weeks intensive course in Modern Industrial Spectroscopy at Chestnut Hill, Boston, Massachusetts, from July 12 to July 23. The course is particularly designed for chemists and physicists from industries in the process of installing spectrographic equipment. Information on the course can be obtained from Professor James J. Devlin, S.J., Physics Department, Boston College, Chestnut Hill 67, Boston, Massachusetts.